PEARSON ALWAYS LEARNING

Final Report
Evaluation of the Chile PSU
22 January 2013

Table of Contents

Table of Tables

Table of Figures

GLOSSARY

2PL	The 2-parameter logistic item response theory model
3PL	The 3-parameter logistic item response theory model
A flag	An ETS-originated term to describe a test item showing little or no differential item functioning according to a set of specific statistical criteria
AERA	American Educational Research Association
anchor set	A set of test items are common across test forms or test administrations and can be used to equate them
APA	American Psychological Association
B flag	An ETS-originated term to describe a test item showing moderate differential item functioning according to a set of specific statistical criteria
bias	A general term to describe a test used unfairly against a particular group of persons for a particular purpose or decision
Bilog	A item response theory software package
C flag	An ETS-originated term to describe a test item showing severe differential item functioning according to a set of specific statistical criteria
CMO	*Contenidos Mínimos Obligatorios*
correction for guessing	A formula used with multiple-choice items for adjusting each test-taker's raw score by estimating how many items the test-taker would have guessed on. Also called **formula scoring**.
CRUCH	*Consejo de Rectores de las Universidades Chilenas*
CSEM	Conditional standard error of measurement
CTA	*Comité Técnico Asesor*
CTT	Classical test theory
DEMRE	*Departamento de Evaluación, Medición y Registro Educacional*
DIF	Differential item functioning
DIFAS	A differential item functioning analysis software package
DTF	Differential test functioning
equating	The statistical adjustment of test scores to achieve comparability across test forms or test administrations
ETS	Educational Testing Service
formula scoring	A formula used with multiple-choice items for adjusting each test-taker's raw score by estimating how many items the test-taker would have guessed on. Also called **correction for guessing**.
ICC	Item characteristic curve
IRT	Item response theory
item bank	A software database for maintaining information related to test items
MINEDUC	*Ministerio de Educación de Chile*

NCME	National Council on Measurement in Education
NEM	*Notas de la Enseñanza Media*
OF	*Objetivos Fundamentales*
PAA	*Prueba de Aptitud Académica*
PCE	*Prueba de Conocimientos Específicos*
postulation score	A score derived from a combination of an applicant's PSU scale score and NEM and used by universities to make admission decisions
PSU	*Prueba de Selección Universitaria*
SAS	A data and statistical analysis software system
SEM	Standard error of measurement
SES	Socio-economic status
SIRPAES	*Sistema de Información de los Resultados de las Pruebas de Admisión a la Educación Superior*
TC	Technical Counterpart
TCC	Test characteristic curve
TIF	Test information function
UCP	*Unidad de Construcción de Pruebas*

Executive Summary

Introduction

International evaluation of an assessment program introduces an additional frame of reference to that already provided by ongoing internal evaluations of a program. Once equipped with this multi-faceted perspective, national program stakeholders can more constructively plan for the future of a particular assessment program.

In June 2011 the Chilean Ministry of Education (MINEDUC) and the Council of Rectors of Chilean Universities (CRUCH) put forth a tender to evaluate more extensively the quality of the entire PSU test battery. MINEDUC and the CRUCH were interested in an evaluation study that covered two main areas. One area was the evaluation of PSU construction processes and items. This area comprised fourteen evaluation objectives ranging from PSU item development practices to its scoring processes. The other area covered the evaluation of the validity of PSU test scores; namely, PSU's internal structure, evidence of its content validity, the change of test score performance, and evidence of predictive validity concerning college outcomes.

In response to this tender, MINEDUC and the CRUCH selected Pearson to provide an evaluation of the PSU that ranged from test construction to the validity of PSU test scores. What follows is an executive summary of the evaluation report that we prepared.

The first section of this executive summary presents an overview of the PSU, explaining its origin, its purpose, the tests it comprises, and how the tests are used in the university selection process. This is followed by a description of the purpose and structure of the evaluation. Next, we provide the key findings of the evaluation. Finally, we present the key recommendations for our evaluation of the PSU.

Overview of the PSU

The Origin of the PSU

The battery of tests that make up the PSU was created by order of the CRUCH in 2001 and is based on the Fundamental Objectives (OF) and the Minimum Obligatory Contents (CMO) of secondary education (*enseñanza media*) prepared by MINEDUC in 1998 (Ministry of Education, 1998). The structure of the PSU was specifically based on the Secondary Education Curriculum Framework (*Marco Curricular de Enseñanza Media*) at the express request of the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior*, (DEMRE, 2010a).

MINEDUC and the CRUCH jointly shaped the PSU, which was developed in response to the need to revise university admission processes in light of Chile's curriculum reforms taking place in the 1990s. MINEDUC and the CRUCH invited a diverse group of organizations and professionals to form a commission to analyze Chile's university admission tests in light of the reform to Chile's national curriculum. The commission formulated the following recommendation to "conduct diagnosis, establish guidelines and propose amendments to the Admission Tests to universities, which will allow it to converge with the new educational goals and objectives raised by the Curriculum of Educational Reform" (DEMRE, 2010a, p. 6).

After meeting for a year, the commission proposed abandoning the idea of testing aptitude on which Chile's university admission tests (*Prueba de Aptitud Académica*-PAA and *Prueba de Conocimientos Específicos*-PCE) had previously been grounded. Instead, the commission

proposed a new configuration for Chile's university admission tests to replace the PAA and PCE with a set of four tests tapping Chile's national curriculum of *enseñanza media* for the following content areas:

- Language and Communication
- Mathematics
- History and Social Sciences
- Science – Biology, Physics, and Chemistry

The purpose of these PSU tests in Chile's university selection process for universities belonging to the CRUCH is captured in the following policy statement:

> The purpose of the admissions process is to select the candidates that apply to be accepted in one of the twenty-five institutions forming part of the CRUCH. The objective of the system is to select those applicants that obtain the best performances in the battery of tests composing the PSU, under the assumption that they represent the best possibilities of successfully complying with the tasks demanded by higher education, for them to enter according to their preference, to one of the institutions forming CRUCH, in the careers they are applying for. Said purpose is achieved by means of the application of educational measurement instruments (PSU), along with the inclusion of the average scores during Secondary Education (NEM). (MINEDUC, 2011, p. 29)

In 2003 the Department of Evaluation, Measurement and Educational Record (DEMRE) at the University of Chile (the group responsible for the development and construction of assessment instruments and measuring skills and abilities of applicants to higher education) took the conclusions of this review and incorporated them into the creation of the PSU, and the tests were first administered for the university selection process in 2004.

Eight private universities were a part of the PSU admissions process with the CRUCH in 2012. Some of these universities were already asking their candidates to present evidence of having taken the PSU in order to apply, but they now follow the full PSU admission process. The eight universities were: U. Diego Portales, U. Mayor, U. Finis Terrae, U. Nacional Andrés Bello, U. Adolfo Ibáñez, U. de los Andes, U. del Desarrollo and U. Alberto Hurtado.

Previous Evaluation of the PSU

In 2004, Educational Testing Service (ETS) conducted an external evaluation of the tests that together form the national admission test battery known as the *Prueba de Selección Universitaria* (PSU). The purpose of the evaluation was to appraise technical adequacy of the PSU Language and Communication and Mathematics tests in the areas of validity, reliability and score use. The evaluation effort covered reviews of PSU test documentation and two meetings with DEMRE staff in Santiago (Educational Testing Service, 2005).

The evaluation effort identified the following strength of PSU test processes:

1. DEMRE staff qualifications and dedication to the PSU testing program

The evaluation effort also identified several areas for improvement of the PSU tests to:

1. Develop documentation of all of the uses of the PSU tests, over and above their use for admission decisions

2. Validate the uses of the PSU tests
3. Develop equating plans for the PSU tests
4. Develop a plan for pre-testing items that allow for comparison across administrations
5. Involve IRT processes for the PSU test construction activities
6. Develop a comprehensive plan for equating PSU test scores across years
7. Develop research on the PSU tests and make public the research reports
8. Develop guidelines for the PSU test score interpretations for relevant audiences
9. Develop plans for introducing DIF analysis for relevant sub-groups

In addition, the evaluation outcomes identified specific areas for improving the technical adequacy of the PSU tests.

1. Adjust the difficulty level of the PSU tests to applicants' level of ability. The PSU Mathematics test was too difficult for the population of applicants. The PSU Language and Communication test, on the other hand, showed adequate difficulty for the population of applicants.
2. Research PSU test bias for all the relevant subpopulations and use the outcomes for improving test construction activities. The evaluation showed the presence of difference on PSU test performance for some of the subpopulations.
3. Expand test score reliability analyses to all the relevant subpopulations (e.g., type of school and region) and develop documentation for reliability process and outcomes for those subgroups.

The Structure of the PSU

The PSU test frameworks came to be adjusted as the system was implemented. The test frameworks referenced domains, defined by Fundamental Objectives (OF) and Minimum Obligatory Contents (CMO), which were formulated as part of Chile's national curriculum for high schools. The adjustment of the test frameworks to Chile's national curriculum took place gradually during the first three years of PSU administration, increasing the coverage of a larger proportion of curriculum content. Revisions of the PSU test frameworks were completed for the 2007 admissions process in Mathematics, History and Social Sciences, and Science (Biology, Physics, and Chemistry) and for the 2009 admissions process in Language and Communication. The theoretical frameworks of the 2007 PSU measure the contents of the General Formation Plan that can be assessed with a paper and pencil multiple-choice test.

Through DEMRE, the PSU has been administered once a year to all applicants seeking admission to any career track in CRUCH universities. The battery of tests is divided into two obligatory tests ("Mathematics" and "Language and Communication") and two optional tests ("Science" and "History and Social Sciences"). One of the two optional tests must be included as a selection factor for each university career.

The Science test consists of a common section that applicants must take along with one of three elective modules in Biology, Physics or Chemistry. Whichever module the student chooses, one single score is assigned to the Science test, which is obtained by using a process that "links" the scores of the common module with the scores of a specific elective module.

For all PSU test batteries, the test questions (also called "items") are *multiple-choice* in which a student selects an answer to a question from five alternative answers of which only one is correct. The total raw scores of each test are calculated by first adding up the number of correct answers and then adjusting the total by subtracting from the number of

correct answers—one fourth of a point for each incorrect one. Items where the answers are omitted are scored as zero. This procedure is intended to adjust the score for the potential effects of guessing.

These corrected scores are statistically transformed to produce a reporting scale that has a mean (or average score) of 500 scale score points and a standard deviation (i.e., a measure of "spread") of 110 scale score points. The extremes of the PSU scale are capped at 150 and 850 scale score points to provide a "floor" and a "ceiling" to the scale. The last step in the process is to smooth the distribution of PSU scale scores to reduce the lumps at the upper end of the distributions.

How the PSU Is Used

The next step in the university selection process is to use the PSU scores to create a *weighted admission score*. The weighting process takes into account weights previously decided by universities for their careers and the weight each university/career gives to each applicant's high school grade point average or NEM, along with the weights for the PSU tests considered in the university selection process. The requirements for each career are reported in the publication *CRUCH Series: Preliminary List of Careers,* halfway through the year. The CRUCH provides norms on the lower and upper bounds of weights for the admission criteria and universities, and their career centers are solely responsible for assigning the specific weights within the range of permissible weights. Selected sets of weights may vary not only across careers and universities but also within a career over time. DEMRE communicates admission results to applicants via the DEMRE portal and to universities utilizing a process defined for such a purpose.

The selection process is a shared responsibility among the CRUCH, universities affiliated to the CRUCH, the PSU Technical Advisory Committee (CTA), and DEMRE. The primary roles of CTA, DEMRE and the *Consejo Directivo para las Pruebas de Selección y Actividades de Admisión* are as follows:

Consejo Directivo para las Pruebas de Selección y Actividades de Admisión (CD)

> The CD is a permanent council whose function it is to ensure the development and management of the selection and admissions system, especially the PSU. (Consejo Directivo, 2010)

Departamento de Evaluación, Medición y Registro Educacional (DEMRE)

> DEMRE is the technical organization of the *Universidad de Chile* responsible for the development and construction of evaluation and measuring instruments of the capacities and skills of the graduates from high school education; for the application of said instruments and for carrying out of a national level inter university selection in an objective, mechanized, public and informed manner. At the same time it is the organization in charge of the administration of the selection system for higher education.
>
> This department, and its predecessor organizations, participated in the creation and administration of the PAA, and while is experts were not involved in the proposed reformulation of the new tests, DEMRE's team has been charged with the construction and administration of the PSU since its first application for the 2004 admissions process. (DEMRE, 2013)

Comité Técnico Asesor (CTA)

> The CTA is an agency of the Council of Rectors of Chilean Universities (CRUCH) whose mission is to assist the Board of Directors (CD) as it coordinates and supervises institutions governing all the dimensions of selection and admission to universities, and to be the intermediary between the CD and the technical teams responsible for the development and implementation of the PSU.
>
> The CTA's responsibilities include proposing initiatives to the CD related aspects of their technical function (data processing, weighting, conversions, etc.). Prior to the beginning of the admissions process each year, the CTA sets the strategy for disseminating information that is useful for applicants and the general public as well as the official version of technical information that is disseminated by the mass media and the associated web pages (e.g., by CRUCH, DEMRE, Ministry of Education, universities, etc.).
>
> The CTA also monitors test development and administration by collaborating with DEMRE to solve technical problems related to the construction of the tests, their application and processing of the results thereof. The CTA also guides the scientific studies concerning the social and academic characteristics of applicants. Finally, the CTA manages the processes and decisions related to operations that deal with issues of proper instrumentation techniques, data processing and analysis, as well as the overall process of selection of students to universities. (Comité Técnico Asesor, 2013.)

Results of university admissions testing also have another use; namely, to provide scholarships and loans to students entering into higher education institutions.

In Chile, financial support is available in the form of scholarships and loans to incoming university students who qualify to receive such awards. There is a wide array of scholarships available to students pursuing post-secondary education. The criteria to assign scholarships are based on applicants' PSU test scores as well as other elements such as past academic performance and socio-economic status. MINEDUC has defined general requirements for applying for the scholarships. More information on the types of scholarships and qualifications are available from www.becasycreditos.cl.

Incentives to higher education organizations come in the form of indirect fiscal contributions. The indirect fiscal contribution is directed to higher education institutions that qualify to receive the monetary incentive for every enrolled student for the top 27,500 PSU scores.

> The Indirect Fiscal Contribution (Aporte Fiscal Indirecto) (AFI) is allocated annually by the State to all Universities, Professional Institutions and Technical Training Centers, recognized by MINEDUC as Higher Education Institutions (Instituciones de Educación Superior) (IES), who accept the highest 27,500 best scores of the students registered for first year studies. (MINEDUC, 2012)

La Junta Nacional de Auxilio Escolar y Becas (JUNAEB) provides financial support to qualified students graduating from high school to take the PSU tests. For example,

> The JUNAEB Scholarship for the PSU is a subsidy aimed at financing the total cost of rendering the University Selection Test (PSU), consisting in CLP $26.000 for 2012, for students from Municipal education and Private Subsidized establishments of the year's graduating class. In special cases, students coming from Private Paid educational establishments may apply if they furnish proof of a socio-economic condition deserving the benefit. (JUNAEB, 2012)

The JUNAEB scholarship has allowed for a segment of the population of high school graduates, which traditionally did not have access to the PSU test, to take part in the PSU test administrations. Such an effort of granting equal access to the PSU test registration has resulted in approximately a 30% increase in student enrollment for the PSU test from the Technical-Professional curricular branch.

The Structure of the Evaluation

The Use of Professional Standards to Guide the Evaluation

The work commissioned by the Ministry of Education was a program evaluation of the testing program. We used three sources of professional standards in guiding our evaluation of the PSU. Our primary reference to professional standards was our use of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Additional reference to standards came in our use of the *International Guidelines for Test Use* (International Test Commission, 2012) and the *Program Evaluation Standards* (Yarbrough, Shulha, Hopson, & Caruthers, 2011).

Each of these professional standards provided a unique view of the responsibilities of test developers and program evaluators. From the *Standards for Educational and Psychological Testing,* we were informed by guidelines in test development, validity, interpretation, and use. From the *International Guidelines for Test Use,* we were informed by guidelines for evaluating and interpreting test scores in a cross-cultural context. Finally, from the *Program Evaluation Standards,* we were informed by guidelines for our responsibilities as evaluators and the importance of recognizing program purpose and stakeholder interests in our evaluative work. Our adherence to the principles set forth in the *Program Evaluation Standards* may be somewhat novel in this context. We believe it is important to recognize that the PSU does not stand alone as a test, but exists within a system of information collecting and decision making that is carried out by MINEDUC, DEMRE and the CRUCH.

Sources of Information for the Evaluation

Our evaluation made use of four information sources:

Formal Documentation. DEMRE, MINEDUC and the Technical Counterpart provided existing formal documentation about the procedures used in developing the PSU and generating PSU scores.

Individual Interviews. Key individuals at DEMRE, MINEDUC and the Technical Counterpart were interviewed in order to clarify documented procedures and to obtain information about procedures for which official documentation was not provided or not available.

PSU Data. MINEDUC granted access to two forms of data: examinee response data and demographic data from PSU administrations, and item-level and form-level psychometric outcomes from the same administrations. Note that although the data

delivery was channeled through MINEDUC, most of the files provided corresponded to the databases supplied by DEMRE or the universities.

Blue Ribbon Panels. Panels including key stakeholders, MINEDUC and the Technical Counterpart were assembled to obtain additional facts about the program including: administration practices, perceptions about certain features of the program, usability of score reports, assessment quality, fairness and consequences.

During February and part of March of 2012, Pearson reviewed the provided documentation and notes taken at the individual interviews in order to develop an understanding of the test development process followed by DEMRE. The goal of this review was to generate an understanding of the methods, steps, tools, software, roles and responsibilities of individuals involved, and the context for which the process existed. A number of questions were generated from this review. Some questions were focused on clarifying the procedures used, while others were about expected process steps that were not described in the documentation.

Additionally, Pearson submitted a list of role descriptions to MINEDUC, which the Ministry then used as the basis for generating a list of possible interview participants. Pearson invited these participants to the meetings scheduled to take place in Santiago from March 27-29 of 2012. The participants were groups of individuals who would take part in interviews regarding DEMRE's process of communicating admission results and PSU content validity.

The Evaluation Framework

A thorough review of any program begins with identifying each evaluation objective characterizing the program. In the context of the evaluation of the PSU tests, there were 18 objectives broken out into three groups. The first group of 10 objectives concerned the development of the PSU tests, while the second and third groups of objectives (four objectives each) examined the test scores arising from the PSU tests and validity studies respectively. Each objective corresponded to an evaluation requirement defined in the PSU evaluation tender.

For every objective, facets were defined based on best practices for large-scale testing of academic achievement. As evaluative facets were defined, relevant characteristics were taken into account through a series of meetings with relevant stakeholders, such as the Technical Counterpart and DEMRE. Major elements of each objective and facet were coded for the professional standards that they addressed using the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

The evaluation objectives were presented to the Technical Counterpart and discussed during the evaluation team's first visit to Chile in January 2012. The goal clarification process took place over three days and provided a means for participants to familiarize themselves with an initial set of evaluation objectives and facets, to propose modifications and, in some cases, to add new elements defining the evaluation facets. A summary of the evaluation objectives is provided in Table 1 and further discussed in the next section.

Table 1: Evaluation Objectives and Number of Facets

Evaluation Objectives	Number of Facets
1.1.a. Item development	7
1.1.b. Item piloting	4
1.1.c. Test construction	6
1.1.d. Item banking	4
1.1.e. Pilot sampling and item selection	3
1.1.f. Pilot vs. operational item performance	4
1.1.g. Exploratory sources of DIF	2
1.1.h. Standardized test scores	2
1.1.i. Reliability and conditional standard error of measurement (CSEM)	2
1.1.j. Recommend a model to derive cut scores	7
1.2. Analyze process used to derive a single score for Science	6
1.3. Evaluate IRT methods to calibrate items and equate scores	7
1.4. Evaluate software for item analysis and item banking	3
1.5. Evaluate score reporting	3
2.1. Internal construct structure	N/A
2.2. Content validity	N/A
2.3. Change in test score performance	N/A
2.4. Prediction of college outcomes	N/A

Note: N/A denotes not applicable.

The objectives having to do with the development of the PSU tests or the use of the PSU (Objectives 1.1.a–1.5) typically required a facet-by-facet evaluation. For these objectives, the PSU processes represented by each facet were evaluated in light of (and coded to) the professional standards from the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). A general evaluative statement and evaluation ratings were provided for each facet, and recommendations were made as to how to improve the specific PSU processes.

The objectives related to the validity of the PSU tests (Objectives 2.1–2.4) took the form of standalone studies. The focus of these studies was on analysis of PSU data that would lead to new information regarding the PSU tests. These studies took the form usually seen in professional journals: an introduction, methodology section, data description, results, and discussion.

Overview of Key Findings and Recommendations by Objective for the PSU Tests

The following two sections summarize the key findings and recommendations of the evaluation report for the PSU tests. The first section deals with the evaluation of the construction and uses of the PSU tests found in Objectives 1.1.a–1.5. In the second, we review the main findings from our validity studies found in Objectives 2.1–2.4.

Within each section, we provide our findings and recommendations on an objective by objective basis. Boldfaced recommendations are considered by the evaluation team to be especially noteworthy with respect to that objective.

Before moving to the specific sets of findings and recommendations, we make note of a very important general recommendation: the need for more extensive and clearer documentation of the processes associated to the construction, application and analysis of the PSU tests. *This need for more detailed documentation is vital given its importance with respect to the improvement of the PSU admissions system.* This recommendation will be mentioned specifically in several of the objectives that follow.

Evaluation Objectives 1.1.a–1.5: The Construction and Uses of the PSU

These objectives examined the framework and specifications used for the item development process (Objective 1.1.a), the piloting of test items (Objective 1.1.b), and the criteria for selecting items for the final test forms (Objective 1.1.c). They also included objectives related to the management of the database containing the PSU test questions and statistics—the *item bank* (Objective 1.1.d), the quality and consistency of results based on test items from the pilot testing and used on the operational PSU tests (Objectives 1.1.e–1.1.f), and statistical analyses of item bias called *differential item functioning* (DIF) (Objective 1.1.g).

The next two objectives examine the procedures used to produce scores on the PSU tests (Objective 1.1.h) and evaluate the procedures used to estimate the reliability of those tests (Objective 1.1.i). The following objective relates to recommending a method for defining cut-off points on the PSU scale in order to provide for social benefits such as scholarships (Objective 1.1.j).

Objective 1.2 analyzed the current use of a single score for the PSU Science considering that this test includes elective modules of items from Biology, Physics or Chemistry. The evaluation of a different approach to developing the PSU tests and scores using the statistical methodology called *item response theory* (IRT) was the subject of Objective 1.3. An examination of the software and processes used for the statistical analysis and item bank of the PSU tests was the focus of Objective 1.4. Finally, Objective 1.5 considered the process of reporting back scores on the PSU tests.

Objective 1.1.a: Quality, security and confidentiality standards regarding the development of items and tests: training of drafters, item drafting, revision, test assembly, printing, distribution and application

Description:

This objective included a review of several aspects of item development and test administration processes used for the PSU, including the:

- Process to develop PSU frameworks and specifications guidelines for developing and writing items (Facet 1)
- Process to select and train PSU item writers (Facet 2)
- Process to commission writing of PSU items (Facet 3)
- Process to review and accept draft PSU items (Facet 4)
- Item authoring tools and item bank (Facet 5)
- Process to distribute and administered PSU tests (Facet 6)
- PSU Scanning and scoring processes (Facet 7)

For this objective, the evaluation team ascertained the principal participants and documented the procedures implemented with an eye toward qualifications of the participants, the possibility for review of the specifications and items developed by another set of experts, the quality of the security and network backups of the system on which the items and test were constructed as well as quality of, and documentation for, the procedures for the distribution and return of materials.

Method:

Information for this objective was obtained by means of interviews with relevant stakeholders from DEMRE on March 19 and 20 of 2012, including the director of DEMRE, the head of the department of test construction, the coordinators of the four subject area test construction committees, the head of the research unit and his team, as well as the heads of the logistics unit, educational records, the computer unit, the admissions process, and the general coordinator. Formal documentation from DEMRE, where available, was also consulted in preparing the evaluation of this objective.

Findings:

The technical teams in charge of writing the specifications have academic training and professional experience that support their engagement in the development of test specifications. The test frameworks and the specifications are the product of a curricular analysis that has also contemplated the demands of higher education in order to pose a serious and objective assessment proposal. However, neither in the framework nor the specifications construction is there evidence of outside DEMRE participation of experts in the review and analysis of the pertinence and relevance of these specifications for the purposes of the test and for the interest of the population under assessment.

Even though the theoretical framework documents and test specifications for each test exist, the depth and detail of the amount of information is not well standardized across the tests and their committees.

The technical team in charge of coordinating the selection and appraisal of the item writers has the adequate academic and technical information to carry out these labors. The

selection process is characterized by being open, which contributes to the transparency of the process and to the variability of the writer teams.

The item writing process seems to be given an adequate time period, which is positive for the process because it ensures that an adequate amount of time for analysis and adjustment can be dedicated to each item. The item developers write the items from the respective institutions where they work, which allows for their availability when developing items and attending item review meetings. The number of items assigned to each item writer is reasonable. However, the review process could be improved if the coordinator of each commission performs an item review before they are taken to the joint review meetings, making use of standardized checklists for compliance with basic item quality criteria. This review would facilitate the feedback to the writers for identifying which item aspects do not comply with the established standard and would optimize the time of the review meetings.

The template used for item writing has key information for characterizing the constructed item, even though it could include more precise information on, for example, the person in the commission who is in charge of performing the first item, etc. This information could later be fed into the item bank and would facilitate subsequent selection processes for test assembly and quality audits.

The existence of a systematic procedure for item review and approval is acknowledged. However, even though the reviewer observations are documented there is no evidence of documentation containing the item quality standards expected for each of the tests. In fact, the rating scales for item approval or rejection used on the different tests are not the same, which is evidence that different criteria are being used in valuing item quality.

In general, the administration and handling of the item bank seems to obey clear criteria and safety protocols that provide reliability to the confidentiality of the items before their application. However, the documentation that describes the process for item entry into the bank could be much more detailed. It is also evident that the documentation does not include a detailed description of the criteria for updating the item bank.

With respect to the safety protocols described by the documentation, it is clear that the restricted access to the physical space of the bank, along with the profile restrictions concerning access to the system, has played an important role with respect to the preservation of item confidentiality. Nevertheless, once again, this positive finding does not mean there is not room for improvement. For example, there is no documentation with respect to the auditing procedures to be implemented periodically in order to detect possible information leak points. Furthermore, although the tools for item *banking* that are reviewed here have been found to be adequate, the item *selection* process, i.e., the decision-making procedures for selecting items reviewed in Objective 1.1.b. is not adequate.

It is acknowledged that adequate safety protocols have been established for handling the examination materials during the printing process, and that the same may be guarantors of the non-disclosure, total or partial, of the test material before an operational test administration. However, the documentation does not provide evidence for the existence of protocols for the control of material in its distribution to the test administration sites.

With respect to the contingency plans for lost or misplaced test booklets, we found this process to be at the level of security required for this type of high-stakes examination. This process is comparable to those found internationally for high-stakes examinations. For example, in the United States, the delivery system of test booklets utilizes a unique

identification number for each booklet for tracking purposes. This numbering process allows for reconciliation of test booklets shipped and returned.

The evaluators recognized serious problems in the use of correction-for-guessing (or formula scoring) adopted for the PSU. Because of the essential role such scores play in reporting of PSU test scores, the international evaluation team regards its use as inadequate because it challenges the validity of PSU scores and PSU field test administration results. For a more detailed evaluation of this policy, see the discussion related to Objective 1.1.h.

Recommendations:

1. **In order to provide for a more rigorous test design, we recommend that the documents state clearly the purposes and uses of test scores and the nature of decisions to be made with the scores (e.g., norm reference/criterion reference), the intended population of test takers, the definition of domain of interest, and the processes outlining the development of test frameworks and test specifications.** Because the PSU is often referred as a battery, it is pertinent to expect a consolidated test specifications document for the tests it comprises. The international evaluation committee would like to stress the importance of developing such documentation with an intended user in mind.
2. In relation to the rigor that the definition and explanation of the dominion of the test must have, the recommendation is to advance studies on the effect that the decision to approach the test with a priority on the CMOs of grades 1 and 2 in high school education may have, as well as determining the effects of the fact that the test may have greater weight for the Scientific–Humanistic curricular branch than for the Technical-Professional one. It would be desirable to foster the materialization of the policies referred to during the interviews with the technical team in the sense of including in the PSU aspects of the Technical–Professional curricular branch, or to study alternatives for the equitable assessment of the populations formed under both curricular branches.
3. **According to the need to include an expert external review of the specifications, it is recommended that the theoretical frameworks and the specifications of the tests be submitted for validation by readers from outside the committee members who do not have the function of constructing items.** This independence would allow for feedback concerning aspects such as the adequacy of the content coverage (axes), as well as the decision concerning the inclusion or noninclusion of CMOs in the same. Part of the framework validation process should be carried out with the participation of experts and with pertinent documentation. The description of the expertise of reviewers is important to collect and report within PSU test framework documentations. This documentation should cover descriptions and examples of materials evaluated, with particular emphasis on test domain, test use, and intended populations of test takers. Finally, we strongly recommend developing, administering, and summarizing participants' answers to a survey after the evaluation exercise and using their recommendations for improving the evaluation process for subsequent years.
4. We recommend including much more explicit documentation regarding the participation of item writers, capturing educational specialty levels, length of teaching experience and region of origin in the country. The use of information technologies could be used for this purpose, enabling the participation at a distance of item writers for whom reaching the capital city of the country is difficult.

5. **Item writer training and certification processes should occur to ensure the quality of the tests and the validity of the evaluation process.** The participants should respond to challenging criteria that formalize the process so that all of the item writers engage in the work under similar conditions. They should have a basic grasp of the discipline in which they are to construct questions; a basic knowledge of the purposes, theoretical framework, and the test specifications; and, finally, the basic technical knowledge concerning the process of test item construction. Each committee allocates a different amount of time to the training process and the work times, and the emphasis on the subjects treated in the induction meetings for item writers are not standardized. In order to ensure an adequate competency level for an item writer, it is necessary to standardize the training process with respect to the objectives, topics, times, materials and the rest of the resources, as well as process control and evaluation mechanisms. It is necessary to give each writer the opportunity to develop his or her item writing skills and to receive timely feedback on the items before that writer begins writing items for the pilot. The Language and Communication committee follows the most formalized item development process. It would be worthwhile to generalize this process to the remaining content areas so as to ensure uniform training of the commission members**.**
6. We also recommend implementing a certification system for item writers completing the formal training process in order to develop a pool of certified item writers. This could be done by DEMRE directly, or an institution that may offer training to writers. In any case, the training process should be designed to include:

    - Verification that the item writers have a grasp of the discipline in which the items shall be developed, emphasizing the assessment content
    - Theoretical training in technical aspects of the test design, such as specification matrices, item construction rules, scales and results reporting, and measurement concepts such as validity and reliability
    - Training in item writing through workshops with detailed feedback concerning the individual mistakes in construction
    - Training in item analysis, including conceptual aspects of psychometric indicators and practical interpretation exercises of same, emphasizing the relationship between item construction characteristics and the indicators obtained
    - Training with respect to the particular purposes of the PSU, its background and uses

    The training shall include participant evaluations to verify a minimum level of understanding of the topics as well as the quality of the items produced.

    The final purpose shall be to grant each participant status as a certified item writer. Also, depending upon the changes introduced in the test or in its development procedures, periodic update processes should be contemplated for the certified item writers. Considering the quantity and characteristics of the topics recommended for item writer certification, we would estimate that a complete course of item writer training such as the one described above would take at least 30 to 40 hours.
7. We recommend performing systematic checks on the assumptions that the item writers are following the principles of confidentiality and copyrights. These checks would entail having senior content staff from DEMRE perform random checks on test content item and art against major sources of copyrighted materials.
8. As stated in Standard 3.7, "[t]he procedures used to develop, review, and try out items, and to select items, from the item pool should be documented." We

recommend clear and transparent documentation of the process for surveying the item bank to identify the quantity of items to be commissioned. Along this line, we would like to recommend specifying more precisely how item writers are to be selected.

9. **We recommend studies that identify the characteristics of the items that can be adapted during item development (such as graphic materials included, letter fonts, diagramming and editing aspects in general, among others) which may make reading of same easier for the population with special disabilities.** Although the procedure established by the UCP to increase the font size and graphics for the visually impaired is relevant and represents an important element in ensuring equity in the assessment process, there is still a need to explore alternative mechanisms, both in the actual construction of testing and in the implementation process to ensure more equitable conditions for all students. For example, what are the local physical conditions of administration or the distance travelled by disabled people to the administration sites? **It is also important to investigate whether the granted accommodations match those used in the classroom. Intended accommodations, when they are not aligned to classroom conditions, have the potential of introducing construct irrelevant variance instead of removing it.**
10. Overall, the tools utilized for item authoring are appropriate for the task at hand. They provide the means and the secure environment required for this type of high-stakes testing development.
11. We recommend increasing the efficiency of the initial draft item review process by involving the senior content staff from DEMRE prior to the full committee review. The purpose of this senior review is to check compliance to item specifications, e.g., the content relevance of items, the appropriateness of the items for different populations, and the application of editorial guidelines.
12. We recommend that DEMRE formalize the process of providing feedback to item writers in a clear and objective way. Such documentation would provide information for analyzing common errors during item writing and therefore guide future training of item writers and reviewers.
13. We recommend that the procedures used to develop, review, and try out items, and to select items, from the item pool should be documented.
14. We recommend using a panel of item reviewers that is independent of the panel of item developers. This item reviewer panel should consist of a group of qualified item writers who did not participate in the development of the particular set of items under review.
15. The Item Bank captures essential characteristics of items within a secure environment. However, given the permanent technological progress in information management matters, it would be useful to periodically implement internal or external auditing systems regarding bank safety control processes, in order to identify possible vulnerable points for item confidentiality, as well as to improve efficiency in the processes of storage, consultation and atomization of the assembly processes.
16. We recommend documentation for the packaging and distribution plans for the field test and operational forms.
17. We recommend clearer protocols for quality control: identity verification, copy control, and the management of chance events (crisis, illness, bad weather, etc.).

18. We recommend cataloging departures from standard administration process so that DEMRE is better equipped when facing circumstances that lead to such departures. We recommend using such a catalog to evaluate test administration process and provide training to staff participating in such processes. The professional development may allow participants in the administration process (e.g., location heads or delegates) to use their experience as PSU administrators and coordinators to report their suggestions to DEMRE on how the test administration process could be improved in the future.
19. Carrying out studies to discard the effect of time and other variables of the application (letter font, test booklet layout, instructions issued to the students, physical conditions of the locations, etc.) which may affect test performance on the part of the students.
20. Even though the scanning process is thorough, the technical processes followed have not been documented nor are there any reports concerning issues that arise during each administration. We recommend that these technical processes be documented in writing and that annual reports, which record the most recent scanning issues and their resolution, be produced.
21. To date, the scanning process that DEMRE has instituted involves both mechanical and manual inspections of multiple markings. The fact that the process involves two levels of resolution on scanning plus a manual check is commendable because it reduces sources of unrelated variance. Nevertheless, though this information is primarily used for the resolution of individual scored item responses, we recommend the use of these erasure analyses at an aggregate level to further support the integrity of the test administration process. Because of the high salience of the PSU, we also recommend further analysis of potential threats to the integrity of test scores arising from unethical behavior (e.g., answer copying, etc.).
22. **In general terms, no studies pointed at supporting the decisions to adjust the raw scores by correcting for guessing. We recommend implementing prospective research studies to document decisions on the use of various studies to support the decisions made. The international evaluation team also recommends a series of retrospective studies to evaluate any potential effects on PSU test scores and item banking field test statistics, for example, of the decisions made in the past due to use of formula scoring.**

Objective 1.1.b: Quality standards of question pretesting

Description:

This objective included a review of the several aspects of the piloting of items for the PSU, including the:

- Design of pilot studies (e.g., specifications, guidelines and criteria) (Facet 1)
- Decision-making process and criteria for selecting items to be piloted (Facet 2)
- Decision-making process and criteria for surveying item bank in preparation to piloting items (Facet 3)
- Process to review performance of piloted items (Facet 4)

For this objective, the evaluation team ascertained the quality of the sampling plans for the piloting, the relevance of stated procedure for the inclusion of items in the pilot, the relative

efficacy of the process for reviewing the item bank for items for piloting, and the criteria for the review of items once they have been piloted.

Method:

Information for this objective was obtained by means of interviews with relevant stakeholders from DEMRE on March 21 of 2012, including the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process. Formal documentation from DEMRE, where available, was also consulted in preparing the evaluation of this objective.

Findings:

The procedure for selecting the sample adheres to accepted sampling criteria, taking into consideration aspects of the strata that are particularly important for PSU testing (dependency, type of curricular branch, etc.). However, there is a lack of a general purpose driving the piloting of the items and psychometric expectations on pilot results. If the purpose of the pilot testing is to gather item data to be further analyzed by groups of reviewers in item data review sessions, the procedures should clearly state boundaries of psychometric performance expected for the items and the nature and representation of review panels. If the expectation is to estimate pilot performance of items to inform test construction without involvement of data review meetings, something which is necessary for a high stake test such as the PSU, there is evidence that this purpose is not fulfilled because the data indicate drastic changes in item properties between pilot and operational administrations. Because the students voluntarily take the pilot tests and with no particular stakes for them, it may be that the motivation level is different with respect to the target population of the test. This disparity in motivation could explain some of the differences in the item statistics presented in Objective 1.1.f.

In general, the item piloting process needs to be better documented with respect to planning the pilot administration and the criteria for defining the population sizes for each pilot test.

As noted above, although the tools for item *banking* that were reviewed in Objective 1.1.a. have been found to be adequate, it does not follow that the item *selection* processes, i.e., the decision-making procedures for selecting items, are adequate. The documentation concerning this is insufficient and during the interviews additional information was not forthcoming. The decisions with respect to the item selection for piloting seem to address a single objective, which is to close the gap between current and expected item counts in the bank. However, this process also needs to reflect the psychometric characteristics of the items expected from piloting. The description of this process does not permit the identification of additional criteria for selecting items for the pilot. These additional considerations include exploring the psychometric effect of different item formats or the statistical behavior of items based on their location within the test booklet, among others.

According to the technical documentation, the process for item banking requires analysis with respect to the preparation of item construction and the piloting that takes place periodically at the beginning of each year. This existing review seems to focus itself fundamentally on the lack of specifications matrix coverage of each test, which constitutes a valid and important criterion. However, it leaves aside the possibility of designing piloting on a scientific basis to study the psychometric effects on the items, item length, booklet editing, etc. This study would enrich decision-making from the design to administration. The item bank review is carried out independently by personnel responsible for the test, and the

documentation does not indicate that standardized criteria are followed to perform such reviews.

The criteria established for most item data review statistics are reasonable. They are in line with what is seen internationally; specifically, the literature and manuals for software commonly used in psychometrics and instrument evaluation. However, the upper limit for a reasonable omission rate (i.e., reasonable <50%) is higher than that seen in other programs in our experience. For example, the number of questions omitted by international educational assessments (such as PISA) is not as high as that reached by the PSU assessments. According to the *PISA 2006 Technical Report* (Organisation for Economic Co-Operation and Development, 2009, p. 219), the weighted average of omitted items was 5.41. In the *PISA 2009 Technical Report* (Organisation for Economic Co-Operation and Development, 2012, p. 200) the average number of omitted items, 4.64, was slighting smaller than in 2006. This suggests that a reasonable upper limit for omissions might be more in the order of 10%.

Recommendations:

23. **It is necessary to establish a sound purpose for the pilot. First, rethink the whole process of the pilot by defining the goals and the use and the procedures to be carried out according to this definition.** For example, develop sample size quotas that take into account expected rates of non-participation—and that those rates might be different for different subjects—so that the goal of 1500 participants is uniformly met. Furthermore, analyze the impact of non-participation rates on representation of major socio-demographic variables. **Next, find socially acceptable ways to increase students' motivation to give their maximum performance on pilot administrations. Finally, identify clearly the quality of the items expected and obtain preliminary values of the parameters that are consistent with the final administration.** From the results of the Objectives 1.1.f. and 1.1.g., the pilot administration has little value beyond the marginal analysis of the quality of the items; hence, the low rating in some respects.
24. We recommend that additional documentation be provided for the following areas: the rationale behind the pilot design; data collection and analysis for the PSU pilot study; and the process for predicting the numbers of items and the number of reading passages to be administered.
25. **Although the statistical criteria for the sampling plan for the pre-test have been documented, e.g., curricular branch and school type, we recommend better articulation of the stratification variables.**
26. We recommend that documentation of the criteria for choosing pilot items be supplemented with a systematic articulation of the reasons for item selection that are determined by the expertise of the participants. In accordance with Standard 3.7, the documentation of these processes would ensure the repeatability of the same even when the expert groups involved in the development vary, thereby increasing the reliability of the process implemented.
27. In this sense, we recommend that greater details be provided, including a statistical rationale, with respect to the placement of common pilot items embedded across more than one pilot form.
28. The recommendation is for the planning of the pilot applications to include clear and intended objectives towards the verification of aspects such as the psychometric effect of using the same items with different item group blocks, or on the effect of

the change in position of an item in different booklets, etc. These studies must be documented and should give feedback to test design.

29. We recommend documenting the purpose of the pilot and the rationale for determining what items are needed according to standardized specifications. The documentation for the pilot must also contain the requirements for sampling and analysis of the items after administration and the elements of CTT or IRT that are considered relevant to the design of the pilot.
30. Perform analysis of the documentation for each cycle of piloting and rate its compliance with the procedures described in the previous section. This analysis should be done after the pilot administration and should be well documented. Checklists can be devised to document fulfillment of the pilot specifications, processes and their stages. Overall a system of quality checks should be put in place to monitor the quality and usefulness of the pilot components and outcomes.
31. We recommend documenting the review process for the item bank survey and the establishment of standardized criteria guiding such processes for all tests, or, in case it is necessary, the justification in order for that process to take place differently for each test. Having manuals to provide the rationale for the pilot analyses ensures that decisions made about aspects of piloting do not neglect the statistical criteria for selecting items.
32. We recommend documenting the plan for communicating the survey results among the functional groups. Pilot studies results should be issued in a systematic way among the teams responsible for the tests. For the piloting to be effective, it should contribute to the improvement of the design, construction, review and assembly of the test.

Objective 1.1.c: Criteria for question selection for the assembly of definitive tests

Description:

This objective included a review of the procedures for the assembly of the PSU, including the:

- Intended and unintended uses and meaning of PSU test scores and intended population of test takers (Facet 1)
- PSU test design and specifications followed when creating an operational form (Facet 2)
- PSU test construction specifications (Facet 3)
- PSU specification matrix orienting the creation of an operational form (Facet 4)
- Process for pulling an operational PSU form and the criteria followed for the creating process (Facet 5)
- Process and criteria for reviewing and approving an operational PSU form (Facet 6)

For this objective, the evaluation team ascertained the historical context for the initial decision to pursue the development of the PSU, its application as a normative test for the admissions process, its purported grounding in the high school curriculum, its correspondence to international standards with respect to its explanation of the proper interpretation of the PSU test scores, the specifications and actual practices followed in the development of the PSU test forms, and whether the tests ultimately constructed can be considered reliable.

Method:

Information for this objective was obtained by means of interviews with relevant stakeholders from DEMRE on March 21 of 2012, including the head of the department of test construction, the coordinators of the four subject area test construction committees, the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process. Formal documentation from DEMRE, where available, was also consulted in preparing the evaluation of this objective.

Findings:

It is acknowledged that the PSU is a relatively new test, taking into account not only its first administration date, but as well the fact that its development has been completed progressively over time (DEMRE, 2010a). For this reason, it is not possible to talk about substantial changes in the meaning and use of the test scores throughout its history. Even so, it would be expected that as of this date there would have been studies available which would detail the perception of the different test users (direct and indirect). Such studies could already be providing information for deciding how to adjust the PSU content, as well as its formal aspects (edition), its conditions for administration and the release and use of the test results.

There are additional uses of the PSU test results that are not intended, such as those when using the SIRPAES report to pass judgment on the quality of schools. As noted in our general description, DEMRE has made some attempts at providing caveats concerning the use of the report, e.g., not drawing conclusions about the quality of education provided at particular institutions from the PSU results. However, it is not enough to put a caveat on a web site; such messages must be embedded in the report itself. The caveats included in the report are insufficient for clearly communicating the intended use of the PSU results. As a result, the likelihood of misusing the PSU results, e.g., disaggregation of scores by school and comparisons with other buildings, is high.

From an international perspective, the evaluation team's experience with the dissemination of university examination results (like SIRPAES) is that such distribution is limited (beyond that to the targeted universities themselves) to the individual applicants and their high school counselors. The primary purpose of this reporting is to look forward to university admission for each student rather than to look back at the quality of the secondary education institution. In the United States, the quality of secondary education is adjudged by statewide assessments specifically designed for that purpose.

With respect to sampling, PSU design team shows the proper consideration of the characteristics of the target population. There is a basic level of psychometric knowledge among the team members and, with respect to item selection, there is support for the statistical targets from classical test theory (CTT) and international standards (levels of acceptance for statistics indicators). That is, DEMRE duly documents CTT statistics gathered used for the construction of the PSU. However, during interviews DEMRE reported that cases exist where some indicators (e.g., average difficulty level or the discrimination index) depart from the desirable criteria. DEMRE's documentation does not account for analyzing these departures.

It is evident that even though there is documentation describing the criteria considered in the test design, it would be desirable to have the support of bibliographic technical references, as well as of studies carried out with the test data, for each one of the decisions

referred to such criteria. A high-stakes test such as PSU also should include measures of precision as part of the test construction criteria. Such criteria, e.g., either a classical or IRT-based conditional standard error of measurement (CSEM), would allow test developers to focus and minimize errors on particular portions of the score scale.

The evaluation team has examined the documented characteristics of the past operational items used during piloting that have been labeled as "anchors." The evaluation team has found them to be below international standards. These items are in no manner whatsoever anchor items. Although the DEMRE's documentation refers to so-called "anchor sets," in practices these item sets are not used for item calibration and score equating. Even if they were used for calibration and equating, their absolute numbers of items in the sets are so low that they would not suffice to accomplish the task in a valid and reliable manner. (See Objective 1.3 for a more extensive discussion of this point.)

The test construction process rests upon the training of the participants, the DEMRE team professionals as well as the rest of the members of the commission. Even if it is true that the good training level of DEMRE professionals supports the assurance of the process quality, it is clear that a test construction manual would be useful, as a technical and instruction-orienting document for those participating in test construction. In that way the standardization in the communication of item acceptance and rejection criteria and of the guidelines and their construction recommendations would be assured. Additionally, a training process including more training time for future developers (such as the one, which according to the documentation, is carried out with respect to the Language and Communication area developers) would ensure the adequate appropriation of the theoretic framework for the other test domains, of their test specifications and of their item construction guidelines among the constructors and would expand the opportunities to see and to analyze item examples and models of the different formats used for those test domains.

According to what was reported during the technical team interviews, the test places greater emphasis upon the Scientific-Humanistic curricular branch of Chile's national high school curriculum than upon the Technical-Professional curricular branch. It should be pointed out that the level of alignment of the matrices with respect to the implemented curriculum taking place in actual classrooms is not known. But, once again, it would have to be verified.

According to the reviewed documentation, the test assembly process uses the item quantities set for each area of the specifications matrix and incorporates items that comply with the statistical criteria established as acceptable. Since the test assemblers are members of the committee who have participated in the design of the test, their informed criterion is ensured in selecting questions that respond to what is intended to be assessed. Additionally, the review of an expert from the university environment as final reviewer of the assembled test adds safety regarding the pertinence of the instrument in a university selection process. The review process could be enriched if it included a final reviewer representing high school education who knows the target population closely (a teacher of this educational level that has not participated in the question construction process to ensure its independence and objectivity) to validate aspects such as question clarity for the students and to question their pertinence with respect to the curriculum carried out in the classroom.

Concerning the assembly system, even though it is partially automated, with a system that pre-selects items in function of given criteria, according to the description found in the technical documentation, a large part of the selection process for assembly is largely

manual, which reduces the efficiency of the process (requires more time and human resources to have an assembled test). A more automated assembly system would reduce the risk of item duplication within a test and the interference of subjective criteria when a choice has to be made of one among many items with similar possibilities of completing an assembly. Such a system would be greatly benefited by the involvement of the IRT framework in lieu of the CTT framework that is currently used for the PSU. An IRT framework would allow targeting the tests to applicants' levels of ability in a systematic way. From its previous evaluation of the PSU, ETS reported a disproportionate difference in difficulty of the PSU test due to the lack of a test construction target for applicants' level of ability (Educational Testing Service, 2005).

Recommendations:

33. **We recommend a better definition of DEMRE's test construction targets, e.g., a tolerance level for conditional standard of error of measurement. The PSU program should identify the portions of the score scale where greater precision is required and construct the test accordingly.**
34. We recommend documentation of the criteria for test construction. This documentation should list:
    a. participants' characteristics and qualifications,
    b. clear definitions of primary and secondary uses of test scores, and
    c. analyses of the consequences on test scores when departing from test construction criteria.
35. **We recommend that anchor items be used for their intended purposes, i.e., to link forms together for the goal of facilitating calibration and equating. The PSU program should also review the criteria for selecting anchor items—including the coverage level of the specifications matrix cells or, at least, the distribution of these items throughout the thematic axes of each test—to reach international standards. We recommend updating DEMRE's anchor set specifications to comply with international standards.**
36. **We recommend providing a manual for test construction.**
37. **We recommend that the test development training process be unified by generating standardized guidelines that are taught to all of them.** These examples ought to illustrate mistakes that should be avoided and aspects that should be considered for achieving compliance with the established acceptance criteria. **The training process should also provide enough time to verify that the new developers comprehend the frameworks and the test specifications before they start the development task properly as such.**
38. Concerning the implementation of new specifications tables, given the 2009 curricular change, we recommend introducing a validation process of the respective specifications tables with teams of high school education and higher education experts (first semester) to add external validity to the process, emphasizing aspects such as pertinence and relevance of the aspects included in such tables.
39. We recommend subjecting the decision to place more emphasis on the Scientific-Humanistic curricular branch than on the Technical-Professional one to outside validation and to include in the theoretic frameworks of the tests the justification for this decision.

40. The test construction process has been well described within the CTT framework. We recommend that the training be formalized and documented, viz., training in test construction tools and processes to develop statistical targets.
41. We recommend considering automating test assembly to avoid the security risks that might arise in the future from the continued physical handling of the booklets, e.g., those that arise from the repetition of questions or from inconsistencies found within the specifications table.
42. It is suggested to include a reviewer of the assembled test coming from the high school education level, contrasting with the reviewer coming from the university level.
43. **We recommend a transition into the IRT framework for test construction. This transition would better position test construction activities to target the PSU tests to the applicants' level of ability in a systematic way. The IRT framework would also provide more precision, and, hence, reliability at points on the PSU scale where the important decisions are made.**
44. We recommend documenting in greater detail the treatment of the drafts or testing material changed during the successive review processes.
45. We recommend that the outside reviews for the operational test review process represent a larger institutional diversity (that is, for not all of them to be exclusively from the Universidad de Chile).
46. We recommend documenting more extensively the procedure to follow when one of the outside reviewers suggests eliminating or replacing an item from a pre-assembled test.
47. The recommendation is to document in a precise way the instructions on test assembly, indicating the ideal distribution of questions in function of the statistical indicators that are taken into account, that is, the maximum and minimum acceptable item number with a certain discrimination level, etc., in order to ensure test comparability between different applications.

Objective 1.1.d: Quality standards in item bank management

Description:

This objective included a review of the standards for item bank management of the PSU, including the:

- Item bank structure (e.g., logical design, platforms, fields, and records) (Facet 1)
- Item bank tools (Facet 2)
- Security access protocols and processes (Facet 3)
- Process flow for updating and adding records to the item bank (Facet 4)

For this objective, the evaluation team investigated the item bank, the Safe Question software used for test construction, the allocation of statistical values in the bank after data analysis, as well as the security and process flow of the groups that have access to the item bank (the Information Technologies Unit, the Studies and Research Unit, and the Test Construction Unit).

Method:

Information for this objective was obtained by means of interviews with relevant stakeholders from DEMRE on March 22 of 2012, including the head of the department of test construction, the coordinators of the four subject area test construction committees, the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the heads of the admissions process and of information. Formal documentation from DEMRE, where available, was also consulted in preparing the evaluation of this objective.

Findings:

The documents reviewed contribute general and apparently complete information on the item bank. How the bank is organized and the interactions between those operating it and the software are clearly understood. However, the information is presented from the perspective of the item bank's software architecture rather than from the perspective of psychometrics, which happens to be important for this process.

In spite of having a clear structure and a powerful database with much information, there is no mention if statistical information related to the bank use is being produced. The psychometric use of this type of information could guide the PSU developments in the immediate future and in the medium term. It is worthwhile to develop reports which inform the designers on test behaviors and not only on the use of items and their statistics.

There is no mention in the documents on the criteria used in updating the item bank beyond the item quantity by region of the specifications table, nor is there mention of the bank capacity.

In the information provided there is some mention of system protection, backup systems for the information and system auditing. However, the information on the data banks has no backup in other external files, increasing the risk of total loss of the information due to an accident. There are no procedures or resources assigned for the protection of the system from possible penetrations due to computer viruses. There is no programming for updating and maintaining the item bank in the sense of including new technologies which would enable its development in relation to PSU.

Recommendations:

48. **Insufficient technical information was found describing the item bank beyond that related to the architectural base. We recommend supplying the information that is missing regarding its modules, their functionality and characteristics.**
49. Even though the aforementioned aspects of the item bank are clearly laid out, we recommend the production of more precise technical characteristics of the test elaboration process; specifically, technical and use manuals need to be generated to facilitate an understanding of what takes place and the real scope and limitations of the bank.
50. Although there is no specific standard for what indicators should be included in an item bank, we recommend additional item use indicators be added to the item bank. For example, knowing the administration history of an item would allow us to calculate its exposure rate.

51. We recommend a technical document describing the total technical features of the software.
52. We recommend a clearer description of the specific rules for how user-level allocations are made.
53. It is necessary to carry out a detailed inspection of the item bank system to determine the updating needs and modifications given the technological developments.
54. With respect to backups to the information systems, we recommend, if under further investigation it is found that the item bank is not supported with redundant systems,
    a. to safeguard against interruptions in service and to maintain the most recent edits, establishing a redundant service (e.g., servers), and
    b. to protect from catastrophic failure, scheduling incremental daily media backups and weekly full media backups of the database.
55. Currently, the item reviews take place based upon the expertise of the participants in the respective sessions (consensus is sought). However, there is no mention of manuals or standards to comply with. **Therefore, we recommend that documenting the item review criteria.**
56. Beyond the fact that the committee members seem to think that an item is good and possesses a certain difficulty, no other statistical or psychometric elements are applied. No statistical information analysis manuals are mentioned. **Therefore, we recommend that committees additionally analyze items with respect to possible discrimination criteria, of the correct option or of the invalid options (distractors), or with an understanding that the items should function in some particular way.**

Objective 1.1.e: Quality of the terms used in the operative applications, considering the indicators used in their selection and considering indicators of item functioning (indicators of the Classical Test Theory, Item Response Theory, and DIF bias analysis) by genre, dependence and educational mode in the experimental sample and in the rendering population

Description:

This objective included a review of the process for analyzing the performance of items selected for and eventually used on the operational PSU test forms, including the:

- Quality criteria for judging items administered operationally (pilot samples of students and population of students) (Facet 1)
- Process for selecting operational items to render test scores (Facet 2)
- Process for reviewing and approving selected operational items (Facet 3)

For this objective, the evaluation team reviewed the Classical Test Theory criteria used and the Item Response Theory difficulty and discrimination criteria consulted during the determination of the set of items to place on the operational PSU test forms. Special attention was given to the criteria applied and practices followed by each of the subject area committees and whether the committees and the outside reviewers had systematically documented how they prioritized the various item indicators.

Method:

Information for this objective was obtained by means of interviews with relevant stakeholders from DEMRE on March 22 of 2012, including the head of the department of test construction, the coordinators of the four subject area test construction committees, the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process. Formal documentation from DEMRE, where available, was also consulted in preparing the evaluation of this objective.

Findings:

In general, DEMRE uses clear criteria on item selection, viz., indicators of the classic test theory and IRT (2 parameters). In almost all of those criteria, the established acceptance criteria correspond to internationally accepted ranges. The exceptions are the IRT difficulty criterion, which spans a range that is larger than commonly accepted, and the omission level criterion, which happens to be quite elevated, even though it is not the same for all tests. The differentiation in the criteria applied for different tests is explained. However, it is not supported by studies evidencing that such differences have no effects upon the assessment process, taking into account the purpose of the test, the assessed population and the object of assessment. In addition, there is no documentation that describes either the basis for the criteria used for item selection or a procedure to be followed for selecting an item when some criteria are met and other criteria are not met.

The decisions on item selection for operational tests are shared among the members of the technical team who, as it has been described, have academic and psychometric qualifications that enable them to perform this labor. The decisions are based upon team review stages and discussions that confer reliability to the process. The procedures for dealing with items showing extreme behaviors are acceptable because such items are, in fact, reviewed by the appropriate DEMRE content committee using relevant criteria.
In general, the process appears to be adequate and meets with minimal expectations.

The teams responsible for item selection for operational tests rely on documentation for the item indicators and characteristics that must be taken into account. However,
DEMRE does not have an established procedure for making item selections when an item complies with some criteria but not others. This allows for greater subjectivity in selecting items and may affect the comparability of forms across administrations.

There is no evidence in the reviewed documentation that the question selection procedures include comparisons between the behavior of the items in pilot and operational applications.

The DEMRE guidelines explicitly allow piloted items to be edited or otherwise changed prior to operational use. This practice contradicts best practices in operational test form development in that items should not be edited or changed unless the items are re-piloted.

Recommendations:

57. We recommend documenting the reason for the criteria transformation in item selection between 2005 and 2011; specifically, documenting the test history in its technical procedures and the reasons for performing changes on the same.
58. We recommend specifying how the IRT indicators are analyzed: ICC and Information Function. Are they the criteria for acceptance or rejection?

59. We recommend documenting the justification for establishing a much wider (from -5 to +5) acceptance range of the difficulty value that the one indicated as acceptable by the IRT literature (from -3 to +3), as well as describing and supporting with greater precision the decision to apply different difficulty criteria (p) for different tests including controls in case they are necessary to prevent such decisions from being counterproductive for the test purposes.
60. We recommend reviewing and reconciling the differing criteria for acceptable IRT discrimination values (i.e., a $\geq$ 0.6 versus a $\geq$ 0.65), given that items selected with that criterion would be cataloged at a low discrimination level, in accordance with the classification table for this indicator.
61. We recommend reviewing the current criterion used for flagging items for high omission rates, applying a standard drawn from the observations and experience of the evaluation team with respect international assessments (i.e., 10% omissions).
62. **We recommend documenting in greater detail the item selection process regarding what steps are entailed in its planning and how specific criteria are applied to each test. Additionally, in cases where there is partial fulfillment of the psychometric criteria on the part of an item, we recommend documenting the rationale that determines which indicator is to have priority in front of the rest.**
63. **We recommend modifying the software used in item construction so that developed items could be loaded to the bank with the history of their modifications and uses in administration.**
64. We recommend documenting the reasons that were considered for establishing some differences in the review of items from the Language committee and the rest of the committees. We also recommend analyzing the possibility of standardizing these proceedings for all tests, in as much as possible. Whenever this is not possible, we recommend that the arguments be documented and the controls be anticipated so that the differences do not affect the results or otherwise be counterproductive with respect to the purpose of the test.
65. We also recommend that over time the choice of participants for the review processes be deliberately made to increase institutional and geographical diversity. Finally, we recommend documenting the policies with respect to contingencies, such as the number of items not approved by the established criteria being higher than is regularly found.
66. **We strongly recommend that piloted items not be edited or otherwise changed prior to operational use unless the items are re-piloted.**

Objective 1.1.f: Degree of consistency between the indicators of item functioning obtained in the application on the experimental sample regarding those obtained in the rendering population

Description:

This objective included an analysis of the factors associated to differential performance of items between their pilot and operational administrations of the PSU.

For the objective, the evaluation team evaluated subpopulation variables associated to the greatest variability in item statistics from pre-test to operational use.

Method:

Information for this objective was developed through an evaluation team analysis of item-level data provided by DEMRE.

Findings:

Based upon results of analyses performed with both pilot and operational item performance across years, there are significant differences in the item performance indicators between the pilot administration and the operational administration.

Considering the results of the classical test theory values as a whole (difficulty, biserial correlation and omission) the differences are large and significant when the item values are analyzed for the set of all of the years in all tests. For example, the category of omissions increases substantially in the final administration; the assumption is that those assessed, upon learning that the final qualification uses a correction formula, omit those responses for which they have a reasonable doubt. This fact affects, in itself, in a great measure the values of all statistic estimations, causing the item values (not only those of the CTT, but also those of the IRT) to be underestimated or overestimated.

The same effect occurs with the IRT values, which are different between both test administrations. This contradicts the theoretical underpinning of IRT that these values are independent from the sample they are obtained from, if they are representative of the same population.

For all years, the lowest association values between the pilot and the final administration correspond to the biserial correlation: specifically, for gender, dependency and branch. It is necessary to note that these discrimination values are affected by the total score obtained from all test items and the strategies used to respond. That is, the discrimination values are affected by the items with technical problems or by how the pilot is managed, or even by the fact that students know about the correction for guessing used to score the test. In this regard it is expected that the values of the biserial correlation between the pilot and the final administration change substantially, which is seen in the tables and figures. It may also be the case that biserial computations can be affected by difficulty levels of items and thus become low as a result of that methodological artifact.

A high level of omission rates and the use of correction for guessing may also have contributed to discrepancies on item performance between pilot and operational administrations. Because the correction for guessing scoring is known to the applicants, the CTT pilot indices may not be reliable approximations of item performance on the operational context.

In summary, the pilot does provide important information about the quality of the items that can be used to make decisions about them in terms of their inclusion or not in the bank for use in any final administration. However, it is not clear that the data can be used as precise values of the different indices studied. Specifically, the changes of the values obtained in the pilot and the final administration are more or less large and significant. These changes occur mainly due to the lack of consideration of all variables (gender, for example) in the sampling of the population for the pilot administration, or, in some cases, the lack of participants forces the sampling plan (the branch, for example) to be reconstructed.

Secondly, the effect of the correction for guessing system of scoring may induce different strategies among students when participating in the pilot and final administrations. The analyses performed on rates of omission showed larger rates of omission for operational administration than for pilot administrations.

Thirdly, the fact that the pilot is a voluntary administration modifies the selected sample.

Recommendations:

67. **We recommend taking steps so that the changes of the values obtained in the pilot and the final administration are closer together.** For example, greater consideration of variables such as gender should be made during the sampling of the population for the pilot administration.
68. **In accordance with the previous recommendation, and other evidence in the evaluation, it is recommended that DEMRE reconsider the use of formula scoring in the context of the PSU. Such formula scoring is based on theoretical assumptions with weak support and international university admissions programs have abandoned its use or are seriously considering removing it from their processes.**
69. **We recommend analyzing the impact of rates of non-participation on intended representation of major socio-demographic variables during the pilot sampling process.**
70. **We recommend redefining the elements of the sample design of the pilot administration, taking into account the purpose of said administration, the purpose of the PSU, the psychometric theory to be used in the item and test analysis and the scoring scale to be used.** This redefinition includes considering other forms of item piloting such as the inclusion of item groups in the operational administration. These groups of items would not be scored or used to obtain the results of those answering them, but that would provide statistics very close to the data from the operational administrations since those being assessed would not know which items are being piloted.
71. Although the sample participating in the pilot is voluntary and there is a commitment from the institutions selected to have their students participate, it is important to analyze impact of non-participation on intended representation of major socio-demographic groups in such a way as to account for possible bias in the results. Once this analysis is preformed, the historical non-participation rates could be accounted for with over-sampling of those groups going forward.

Objective 1.1.g: Exploration of variables associated to DIF, in case it is present

Description:

This objective included an exploration of the variables associated to DIF in the PSU.

For this objective, the evaluation team provided (1) expert analyses of DIF processes documented in DEMRE reports and clarified during interviews of DEMRE staff, (2) an analytical inspection of variables related with DIF, covering statistical modeling of DIF results with relevant item level information and (3) an empirical demonstration of DIF computation with PSU data from the 2012 admissions process.

Method:

Information for this objective developed through the evaluation teams analysis of data provided by DEMRE and clarified through interview questions.

Findings:

Differential item functioning (DIF) refers to ascertaining the relative difficulty of a test question for one group of test-takers versus another group when matching test-takers within those the groups that have the same level of ability. This statistical process allows test developers to identify if items as potentially biased. Such items are then reviewed to determine whether those differential effects are irrelevant to the construct targeted by the test. The DEMRE documentation presents how it performs DIF studies, how it processes the data for the analyses and how the results are summarized. The documentation also presents criteria used in classifying items with DIF.

There are a few aspects of DEMRE's approach to DIF analysis that have drawn the attention of the evaluation team. One aspect is DEMRE's decision to dismiss information concerning the high prevalence of DIF rates for pilot administrations. High rates of pilot DIF manifest significant problems with pilot conditions (e.g., self-selection, differential motivation rates, and high rates of omissions and the representativeness of the pilot sample). These conditions can introduce an element of bias into the total test score and thus affect its use as a matching variable for comparing reference and focal groups in a DIF analysis.

The second aspect of DEMRE's DIF procedures that raises a concern is its decision to emphasize operational DIF results over the pilot DIF results. The evaluation team considers these decisions problematic because DIF analyses and outcomes are important for assessing fairness proactively during the piloting of items. Addressing DIF through operational results is risky because best practices call for examining the items for bias *before* they are presented to students during operational administration. In an ideal world, quality control processes are set up to detect anomalous items before they become operational. In applied scenarios, removing items showing DIF requires human expert interpretation of construct irrelevant sources behind the statistical flags.

A third aspect is the lack of a policy guideline directing selection of reference and focal groups for DIF analyses. The fact that DIF is only calculated for gender and type of dependency is also troublesome. Internationally, the concept of protected classes of test takers has influenced the practice of defining the groups for DIF analyses. Recently the concept was broadened to include variables to better understand factors behind DIF (e.g., curriculum exposure). DEMRE's analysis should also include such factors as socio-economic status, high school curricular branch and region. Also, we would like to recommend exerting caution on the reliance on multiple comparisons (e.g., Private vs. Municipal, Private vs. Subsidized and Municipal vs. Subsidized) with no particular hypothesis to check. It is obvious to expect that multiple comparisons could have an effect on the Type I error rate and affect the efficiency of the process. DEMRE's documentation does not provide a rationale for using multiple comparisons. DIF flags are not necessarily indicative of bias and DIF analyses should not be carried out mechanically.

A final aspect of DEMRE's DIF processes that raises concern is its decision to use more than one approach to explore DIF. This approach is reminiscent of the man with two watches who didn't know what the exact time was. When used either individually or conjoined, the approaches aim at detecting DIF relatively to the set of test items, with different levels of statistical error (Type I and Type II) and statistical power. DEMRE's documentation does not

appear to contain any rationale for using more than one method of DIF analysis or any statement about the relatively superiority of the approaches.

The DIF analysis is a task that goes beyond the processing of data and the use of standard criteria on those data. The development of procedures that enable the detection of plausible explanations for the presence of DIF should be added to DEMRE repertoire. The evaluation team's analyses of archival PSU DIF data showed a straightforward way to explore for potential sources associated with DIF using logistic regression. This type of analysis could be expanded to accommodate other item attributes, such as the use of words, presence or absence of art and the nature of distractors.

The demonstration the evaluation team provided exemplifies DIF computation for a broader set of demographic or other grouping variables. The follow-up to this demonstration would involve groups of qualified content specialists and teachers in meetings in which items with DIF flags are further analyzed. The outcomes of those meetings are important because they expand the understanding of the factors affecting the quality of items, which can improve item development training and inform item development specifications in the future.

For the pilot, as long the statistical DIF for an item has been analyzed by a bias review committee and these experts have found that the flag was not pointing to any meaningful bias, then the item should be allowed for use if needed to fill content gaps on the test.

For the operational test, if an item is flagged for DIF, it does not necessarily mean the item was biased. A review is needed to determine why the item was flagged. For example, something not found previously (e.g. double keys) and fundamentally wrong (formatting, printing error) could trigger the DIF flag. Omission rates can also muddle the information.

This demonstration found, among other results, that:

- Most of the items showed negligible DIF (A), with very few items showing weak or strong DIF (B or C)
- PSU Science (common and elective portions) showed larger number of DIF C flags than PSU Mathematics, Language and Social Studies
- The Gender variable showed the largest number of DIF C flags for a given Common portion of the Science test (six favoring Males and three favoring Females)
- The variables SES, Region, Curricular Branch, and Modality showed far fewer DIF C flags, with the greatest number of DIF C flags appearing on the Chemistry test for the Scientific-Humanistic versus Technical-Professional curricular branch comparison (three favoring Scientific-Humanistic and four favoring Technical-Professional)

Recommendations:

72. **We recommend evaluating the significance of pilot DIF results as part of the data review processes and prior to banking the items. Items with DIF C flags should be scrutinized for potential bias by data review panels.** Once the items have been analyzed, a record of the decisions reached in data review should be added to the associated item documentation, noting the decision to use or not use the item for operational administration.
73. **We recommend expanding DIF analyses to relevant sub-groups that historically have not been part of DEMRE DIF analyses.** At a minimum, DIF

analyses should be expanded to the following subgroups: region, socio-economic status and curricular branch.

74. **We recommend setting a policy for defining reference groups.** The process currently followed involves multiple comparisons among categories of the sub-group variable which is not only inefficient but also increases the type I error rate for DIF results.
75. **We recommend choosing the Mantel-Haenszel DIF method instead of using multiple DIF methods.** The use of Mantel-Haenszel Chi-squared method is well documented and allows for the use of ETS DIF classification rules. If for any reason a backup method is needed, the evaluation team recommends the logistic regression method. The reliance on a single process should be clearly stated in the documentation. The use of multiple methods becomes problematic because different methods have different Type I error rates.
76. **Once DEMRE picks a single method for calculating DIF, it should involve content experts to examine those items that have been flagged for DIF. We recommend that DEMRE create criteria for invalidating pilot items with DIF outcomes, such as C flags. The process should differentiate between statistical flagging of DIF and content sources of DIF. The international evaluation team strongly recommends avoiding the use of multiple DIF methods.**
77. The evaluation team recommends taking into consideration policy and practical limitations when choosing focal and reference groups for DIF analysis.
78. The PSU program should investigate sources of DIF and use results to fine-tune their item development practices, test construction models, and test scoring process. Analytical approaches can be used to gain understanding on variables that relate to DIF flags.
79. **The PSU program should complement the data obtained with DIF detection analysis with the participation of content experts and educators.**

Objective 1.1.h: Analysis of procedures for the calculation of standardized scores, score transformation in relation to the original distributions

Description:

This objective included an analysis of the standardized scores of the PSU. As part of this objective, the evaluation team computed PSU corrected raw scores, PSU scale scores, and PSU smoothed scale scores so as to compare PSU standardized scores and score transformation in relation to the original distribution of corrected raw scores.

Method:

Information for this objective was obtained by means of analyses performed by the evaluation team and through interviews with relevant stakeholders from DEMRE on March 23 of 2012, including the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process.

Findings:

The international evaluation team recognizes efforts made to provide PSU standardized test scores and their corresponding computational procedures. Collectively, the steps followed for computing PSU scale scores resemble the canonical structure of processes for computing scale scores in norm-reference contexts. The steps respond to admission policy mandated by the CRUCH and DEMRE's role as a data processing entity for the postulation scores. As mandated by the CRUCH, PSU test scores are used to select students for university studies. The PSU test scores are used to rank-order applicants seeking admission into college level studies.

The documentation of PSU processes to develop scale scores provides a good starting point but this documentation needs more elaboration and inclusion of supporting evidence for: (1) correcting for guessing, (2) the rationale for choosing the mean and standard deviation of the PSU scale scores, (3) the decisions for truncating the PSU scale scores, and (4) the maintenance of PSU scale.

The evaluators recognized serious issues in the use of formula scoring to correct for guessing on the PSU. Due to the essential role such scores play in reporting of PSU test scores, the international evaluation team regards its use as problematic because it challenges the validity of PSU scores and PSU pilot test administration results. This can be remedied in future years by adopting item response theory models that account for guessing. The contribution of the corrected scores to the improvement of PSU test score reliability and the predictive validity of scores, and public opinion is not documented.

The precision of the scale in which results are reported is not estimated. The test reliability and standard error of measurement is estimated from raw scores utilizing classical test theory. Precision of scale scores both typical and conditional is not part of DEMRE processes. This is a serious limitation considering decisions are made on the scale score metric. Conditional standard errors of scale scores involved in the university admissions decisions should be calculated and communicated to PSU audiences.

The greatest concern on the development of the PSU scale is the lack of a mechanism put forward for maintaining the PSU scale across years. As mentioned before, in Chile, PSU test performance is reported with the PSU scale but no equating takes place to maintain the scale across admission years. This is a serious issue that must be attended to ensure Chile university admissions test is fair to test takers and ensure valid comparisons of test scores across test administrations. For more information on the evaluation of equating of the PSU, please see Objective 1.3 of this report.

Finally, from the interviews, the evaluation team learned about high level features of the process to standardize the high school grade point average (*Notas de la Enseñanza Media* , or NEM). NEM is computed by averaging final grades attained in high school studies. The grading system ranges from a minimum score of one point to a maximum score of seven points. However, DEMRE was not able to deepen our understanding by providing more information on the structure of the NEM and the process to develop NEM norms because of the lack of documentation existing on those subjects. Due to the role NEM plays in the computation of the postulation scores, it is troublesome not knowing the psychometric properties of the NEM scores. Of special significance is the comparability of meaning of NEM scores across multiple subpopulations. NEM scores are based on grading practices that may or may not be comparable across schools (Private, Subsidized and Municipal) and curricular branches (Scientific-Humanistic and Technical-Professional), for example.

The evaluation team also recognizes serious problems in the documentation and technical adequacy of scales, which has caused the team to disapprove the associated processes for scale development. Most of the professional standards listed for evaluating this objective were not fulfilled (professional standards 2.2, 2.14, 4.2, 4.5, 4.6, 4.8). These deficiencies can be improved in the near future.

Neither the psychometric properties nor a proposed method of interpretation of the postulation score have not been documented. While admission criteria stipulates use of PSU test scores and NEM scores (along a set of weights and policy considerations) as building blocks of the postulation score, little is known about the psychometric characteristics of the composite postulation score such as its scale mean and unit of dispersion. While a normative meaning is attached to the individual PSU test scores, evidence for the meaning attached to the postulation score has not been defined.

Another main problem we detected is the lack of information on measurement precision of the postulation scores. When ranking postulation scores, numerical differences between postulation scores are of no consideration despite their potential lack of practical significance. For example, a score of 672.15 points ranks above a score of 672.13 points; nevertheless, the two scores show differences at the second decimal place. Without measures of score precision available to understand differences in the scores, it is plausible that differences of the size of a few decimal points could be interpreted and used to make decisions on who gets admitted and who does not. In summary, a lack of conditional standard error of measurement for postulation scores is a major drawback that needs to be addressed in the near future. More discussion on needs to compute measures of accuracy and precision is included in the section covering Objective 1.1.i (reliability and conditional standard error of measurement).

The evaluators found some level of documentation of the descriptions of the processes to derive PSU scale scores for individual PSU tests. The reviewed documentation and information from interviews helped in understanding the process to develop the scale and the process to assign meaning to scale points. For example, while documentation summarized PSU scale characteristics, interviews helped the evaluations obtain fine-grain information on the PSU scale. DEMRE documentation needs to be expanded to cover description of measurement precision (e.g., conditional standard error of measurement) for individual PSU tests, which to the present date has been neither computed nor reported to users. It is also recommended that summaries of the limitations of derived scale scores be provided, when applicable. For example, we found the process for manually smoothing the upper tails of the distributions (1%) problematic for the following reasons. The processes followed by DEMRE depend on human judgment and lack of quality control checks. In addition, DEMRE has not produced evidence on the effects of the smoothing on bias reduction.

The evaluators found little to no documentation of the processes for deriving norms for NEM. During interviews with DEMRE staff, international evaluators uncovered the existence and use of NEM norms. The interviews with relevant stakeholders made evident that NEM norms have been used since 2003. From the interview it was also learned that there are three sets of NEM norms: (1) Scientific-Humanistic (morning attendance), (2) Scientific-Humanistic (afternoon attendance), and (3) Technical-Professional, respectively. Throughout the interviews, DEMRE staff expressed their lack of knowledge of the conditions under which the norming was performed and commented on the lack of availability of technical reports on the norming studies. At the time of the interviews, DEMRE was not able to deepen our understanding by providing more information on the structure of the NEM and the process to develop NEM norms because of the lack of documentation. Because of

the role that NEM plays as part of the computation of the postulation scores, the lack of information on NEM norms and the psychometric properties of the NEM scores results are troublesome. The lack of evidence supporting comparability of classroom grading practices is another concern. NEM scores are based on grading practices that may or may not be comparable across schools (Private, Subsidized and Municipal) and curricular branches (Scientific-Humanistic and Technical-Professional), for example.

Recommendations:

80. In Chile, the contribution made by the correction for guessing (formula scoring) process to the improvement of PSU test score reliability, PSU predictive validity of scores, and PSU public opinion is not documented. Internationally, the use of the correction for guessing faces challenges in these areas as presented in the summative evaluation of this facet. Additionally, when omissions are considered as not reached, the corrected by guessing scores vary for students with the same number correct score with different omission rates. The correction for guessing may lead the students to use a strategy for approaching the test which does not have to do with their knowledge, thereby reducing the accuracy of the prediction of university performance. **Finally, in light of international standing of the correction for guessing on university admissions programs, the international evaluation team recommends considering abandoning the practice of correcting for guessing for future administrations.**
81. **We recommend considering item response theory as an alternative approach to deal with applicant's guessing behavior.** Current PSU scoring approaches use correction for guessing processes that have been found with severe limitations in the literature. As mentioned above, the process adds layers of complexity to multiple aspects of the processes such as when calibrating pilot testing responses, computing item statistics and test score statistics. The item response theory framework brings build in features to account for the amount of guessing (i.e., pseudo-guessing) present in test taker's item responses. **We also recommend that in preparation to transition out from the correction for guessing (formula scoring) context, if decided, DEMRE prepares and submits a transition plan to an external expert group of reviewers.** The plan among other aspects should involve risk analyses and feasibility analyses and a time frame to introduce the necessary changes on critical processes of the university admissions testing program such as PSU test construction, PSU item banking, PSU pilot, PSU scale maintenance, PSU validity and reliability, PSU score reports. **The international evaluation team also recommends a series of retrospective studies to evaluate any potential effects on historical trends of PSU test scores and item banking field test statistics, for example, of the decisions made in the past due to use of formula scoring.**
82. Because of the inclusion of the standardized high school grade point average (NEM) into the postulation score, the evaluation team recommends the evaluation of their normative data. Conversion tables for NEM were first used in the 2003 admission process. Because NEM is one of the two elements defining the postulation score for most of careers with the CRUCH and the eight affiliated private universities, we recommend deepen the information available on the structure of NEM and the process to compute it. Along these lines we recommend studying the validity of the inferences drawn from the set of normative data that is almost ten years old and that is based on a national curriculum that has been almost replaced by the current national curriculum. Along these lines of research and documentation, the international evaluation team recommends studying the validity of standardized NEM

scores. Of special significance is comparability of meaning of NEM test scores. NEM scores are based on grading practices of unknown generalizability across type of schools (public, private, and subsidized) and curricular branches (Scientific-Humanistic and Technical-Professional). The international evaluation team proposes studying the possibility of replacing NEM with standardized measures of high school academic performance such as scores from nationally administered tests. It is of crucial to properly balance the PSU test frameworks and their reference to Chile's national curriculum to avoid over-emphasizing measures of high school academic performance while sacrificing measures of general scholastic aptitudes.

83. **We recommend completely revising the postulation score system using the perspective of composite scores.** Procedures for computing composite scores and their associated scale scores are well documented in the literature, and there are diverse methods available to practitioners (Feldt & Brennan, 1989; Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). These efforts should be instituted to provide precise information at those regions of the postulation score scale where important decisions are made (e.g., admission decisions), while acknowledging for the differential composite weights used across universities and their careers.
84. **We recommend introducing the measurement precision of reported standardized scores for individual tests into the PSU test score and postulation score system.** The PSU program should develop guidelines on intended uses and interpretation of PSU scale scores and standardized scores with an emphasis on delineating the limitations of the use and interpretation of derived scores. The efforts should keep in perspective the differential composite weights used across universities and their careers.
85. We recommend providing technical documentation of the norming of NEM scores with an emphasis on descriptions of the intended PSU test populations, sampling procedures, participation rates, weighting approaches (if used), testing dates, and descriptive information of background variables.
86. **We recommend maintaining a research agenda to study year-to-year stability of primary and secondary PSU scales.**

Objective 1.1.i: Reliability (CTT) and precision (IRT), including the information function, of the different instruments forming part of the PSU test battery - Standard error analysis of conditional measurement for the different score distributions sections, placing special emphasis on the cut off scores for social benefits

Description:

This objective included a review of the process for estimating PSU test score reliability from CTT and IRT frameworks and for computing conditional standard errors for PSU scale scores. The various aspects investigated included the effect of the policy of correction for guessing as well as the significance of the reliability of the PSU test scores when used for admissions and for scholarships.

For this objective, the evaluation team also provided a demonstration of the computation of the conditional standard error of measurement of PSU test scores under the item response theory framework (IRT). The standard error of measurement conditioned at a given level of proficiency is known as the conditional standard error of measurement (CSEM). The CSEM estimates the amount of measurement error across the test scale. Test developers use the

CSEM to target higher level of measurement precision at specific regions where the most important educational decisions occur. Failure to meet this goal indicates that the tests have lower levels of precisions at the targeted regions of the scale. To date, CSEM analyses have not been a part of DEMRE's psychometric processes. In the future this type of analyses should be added to the PSU testing program.

Method:

Information for this objective was obtained by means of analyses performed by the evaluation team and through interviews with relevant stakeholders from DEMRE on March 23 of 2012, including the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process. The evaluation team also read the PSU reliability report (DEMRE, 2010b).

This objective also provides a demonstration of the computation of conditional standard error of measurement (CSEM) of PSU test scores under item response theory framework (IRT). DEMRE does not compute nor report the above measure of accuracy as part of their processing of admission test data. DEMRE provides reliability and standard error of measurement from classical test theory. The demonstration used data from the 2012 PSU admissions process.

Findings:

It is a tenet of sound test development and use to document reliability, standard error of measurement, conditional standard error of measurement at each score and their combination into a single composite score. When evaluating reliability estimates of PSU test scores, it is important to consider the use and interpretation of PSU test scores. Certain score uses require greater confidence in the accuracy of the test than other score uses. For example, in Chile, important decisions are made with PSU test scores such as granting university admissions and granting scholarships. If these decisions are to be properly executed, they must attend to the characteristics of university admissions and scholarship granting process. Although there is a unique set of cut scores orienting decisions along the two venues, most decisions tend to be made on the test scale score region above the center of the PSU scale (500 points). When cut scores are used, the amount of information the test produces should be somewhat maximized at those scores, particularly when high-stakes decisions are being made.

The reliance on internal consistency measures (e.g. coefficient alpha) and classical test theory standard error of measurement provides a partial coverage of what is expected for a high-stakes test in terms of international standards.

The process to estimate reliability focuses on reporting estimates for individual PSU tests; nonetheless, admissions decisions are made with composite scores weighting PSU individual test scores and the high school grade point average. DEMRE's reliability report (DEMRE, 2010b) does not provide a rationale for skipping the reporting of reliability of the composite university admissions score, which is ultimately the piece of information upon which university admissions decisions are made. In addition to the lack of an estimate of reliability of the university admission composite, there are no pieces of information on the measurement error for the composite score or on the confidence bands around the reported associated percentiles.

The process to estimate reliability involves typical formulation of coefficient alpha for number-right multiple-choice test. In Chile, PSU test scores are produced with a formula

scoring that (1) removes from a correct multiple-choice response one fourth of a point from every multiple-choice wrong response and (2) leaves unaffected the number right score when applicants omit the item. Although DEMRE relies on the above "corrected by guessing" observed score when computing PSU test scores, the computed PSU reliability coefficients are based on PSU raw scores. Interestingly, raw scores do not account for guessing correction.

The scope of the PSU reliability estimates that we found is limited when providing a rationale for relying on just number correct coefficient alpha and ignoring estimates of classification consistency/accuracy estimates. When continuous scores are interpreted with respect to one or more cut scores, the coefficient alpha and the standard error of measurement may produce information that may be unrelated to the following question: "How consistent is pass/fail classification?" Since the primary use of the PSU test scores is to screen between applicants to university careers reaching the admissions score and applicants not reaching the admissions score, accuracy of classifications is an important piece of information that ought to be included as part of the report, if the audience for that report is to understand the degree of decision consistency achieved. Such pass/fail decisions are better informed with classification consistency and classification accuracy approaches which are standard psychometric practices involving a single test form.

National and international standards recommend using the standard error of measurement as a gauge to compare groups instead of simply comparisons of reliability estimates. It is well known that reliability estimates are group dependent whereas measurement error is not. The reliability report provides information about the amount of measurement error for typical applicants (standard error of measurement). However, as far as the measures themselves are concerned, it would be better to consider measurement precision. Technically, this is the inverse of the error variance of individual measures. In classical test theory, standard error of measurement assumes that error is the same over the entire assessment scale. It is more realistic to assume that standard errors are smaller at proficiency levels where large numbers of items are concentrated. This implies that precision is concentrated at the middle of a proficiency distribution and lower in the tails of that distribution where relatively few items are found. When standard errors are plotted in relation to proficiency, this produces a U-shaped curve.

For university admissions examinations, it is important that the center of this U-shaped curve is positioned over the range in the proficiency distribution where admissions decisions are likely to be made. One can think of this position as a certain range of percentiles within the examinee population. Adding historical information on the percentages of examinees admitted to careers to the measures of PSU test score precision, universities and their careers could include additional information to orient their selection decisions.

Because of the importance of the postulation score on the admission process, the fact that NEM scale score and PSU scale score differ in their dispersion creates concern. With NEM reported on a scale with narrower dispersion (i.e., a standard deviation of 100), variance of NEM in the postulation score would be small and this scaling decision would enter into the variance of the postulation score and its reliability. Postulation scores are weighted composites involving PSU test scores and NEM and adopted set of weights. The variance of postulation scores is a function of (1) the squared weighted variance of PSU test scores and NEM scores and (2) the weighted covariance of the PSU test scores and NEM scores. The international evaluation team recommends addressing the scaling of NEM for future administrations and when computing current norms for such variable. A sensible scaling decision would be to reset NEM scale to use that which is similar to the PSU scale as part of NEM norming activities.

Recommendations:

87. The PSU reliability report (DEMRE, 2010b) is limited when providing justification of the approaches to estimate reliability that have been used. In the analyses reported, the coefficient alpha was implemented to estimate reliability. Coefficient alpha shows specific sources of measurement error relevant for some type of decisions. Specifically, coefficient alpha shows measurement error due to sampling error associated to items.
88. Discussion of plausible systematic sources of errors on PSU scores is also absent from the PSU reliability report. The report is lacking discussion of effects and treatment to accommodate correction for guessing and omission. The effects of these two conditions deserve more study. Similar challenges were noted for the estimation of the standard error of measurement, which relied on the standard deviation of uncorrected raw test scores.
89. **The PSU reliability report is limited because it does not provide information on conditional standard errors of measurement or on the rationale for establishing the acceptable size of such errors for the intended primary (placement) and other (scholarship) uses of PSU test scores.** As shown in our demonstration, item response theory provides a framework for examining in a valuable way the conditional standard errors of measurement. **We recommend that these analyses to be added in the future to the PSU program.**
90. **The PSU reliability report is limited when providing descriptions of the amount of measurement error at critical regions of the PSU scale utilized to make high-stakes decisions (e.g., accepted/rejected admission and granted/not granted scholarship).** Measures of decision consistency/accuracy are import*ant pieces of information currently absent from the* estimate of PSU score reliability and precision. **Additionally, the existing processes do not explain the precision of PSU scores for primary and other decisions (e.g., university admission and granting scholarships, respectively). The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) advises reporting precision of the scores on the scale from which decisions are made.**
91. We recommend addressing the scaling of NEM for future administrations and when computing current norms for such variable. A sensible scaling decision would be to reset NEM scale to use that which is similar to the PSU scale as part of NEM norming activities.

Objective 1.1.j: Propose a model for dealing with cut off points for social benefits, from the perspective of the Classical Test Theory (CTT) as well as from the Item Response Theory (IRT)

Description:

This objective included proposing a model to derive cut scores of the PSU for social benefits.

Method:

Based on its extensive experience with the set of cut scores on high-stakes educational assessments, the evaluation team reviewed the current circumstances in which social benefits are allocated in Chile with respect to the PSU and thereby made a recommendation

of a method (the Hofstee method) that could be pursued by the parties responsible for such distributions of social benefits in Chile.

Findings:

There are no specific findings for Objective 1.1.j, only the recommendations below.

Recommendations:

92. **With respect to proposing an approach for defining cut points on the PSU scale for granting social benefits in forms of scholarships, we recommend an approach that considers both PSU domain mastery and social consequences.** The former aims to identify the level of knowledge, skills, and abilities defined by a panel of university professors and MINEDUC policymakers and to translate that definition into a PSU admission score. The latter takes into account policy considerations, social consequences and historical data to fine-tune the cut score. A focus on social consequences would entail MINEDUC policymakers convening a panel to use such information as the historical data on the number of scholarships available each year, the number of students receiving scholarships, their performance, their attrition rate, and their graduation rate.
93. **The specific method that we recommend for setting the cut score for scholarship purposes is the Hofstee method.** The Hofstee method is an example of a compromise standard setting method. Hofstee (1983) coined this term when he developed his method to capture the dual nature of standard settings. That is, even criterion-referenced standard setting judgments are tempered by norm-referenced expectations. His method makes explicit use of both criterion- and norm-referenced information to derive cut scores. The term "compromise" does not connote the lessening or dilution of either of the two criteria of judgment; rather, it signifies the integration of the two in a manner that is more multifaceted than either criterion applied individually. This is important for awarding scholarships because there are finite resources to distribute and norm-referenced selectivity targets those resources to students in a manner that should be considered along with criterion-referenced subject area mastery. Although the approach is not exempt from criticism, as with any other standard setting process, we believe that setting a cut score for granting scholarships requires a blend of considerations: some of them scholastic, some monetary, and some social. The virtue of the Hofstee approach relies on its capacity for bringing all these perspectives into account to facilitate discussions and compromising decisions within a reasonable time frame. The approach is flexible enough to be used with the current approach to scaling the PSU (i.e., using classical test theory) as well as any foreseeable changes to the system (i.e., using item response theory).

Objective 1.2: Analysis of the adequacy of a single score in the Science test and of the procedures to calculate said score, considering that this test includes elective blocks of Biology, Physics and Chemistry

Description:

This objective included an analysis of the process used to derive a single score for Science of the PSU. For this objective, the evaluation team provided an analysis of the pertinence of the single score for PSU Science. The evaluation team performed demonstration analyses to address two fundamental questions: (1) How reasonable is the reporting of a single score?

and (2) What alternatives exist? The evaluation team also discussed the issue of dimensionality of test scores.

Method:

The evaluation team read the Technical Advisory Committee documentation of PSU Science "equating" process and met with DEMRE staff to obtain additional information on the process for calculating the single Science score. The interviews with the relevant stakeholders from DEMRE occurred on March 26 of 2012, and included the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process.

Findings:

The evaluation team considers the reporting of a single PSU Science score to be untenable because it relies on a questionable assumption of equivalence (e.g., meaning) of the part-test scores (Biology, Physics and Chemistry).

The process that is being followed to render a single Science score is not equating in the strict sense defined by Kolen and Brennan (2004) because the examinees take tests with different content based on the optional sections (alternative modules). Because these alternative modules bring content differences, scores for students taking the different optional modules cannot be considered to be equated. However, such scores can be referred to as "linked," and the process followed referred to as "linking." Terminology associated with the single score in Science needs to be changed from "equating" to "linking."

The process currently used to develop a score for each examinee involves linking the score on each optional section to the score on the common portion using a nonlinear regression method. The single score is the sum of the score on the common portion and the linked score on the optional section. Separate nonlinear regressions are used for the three optional sections to derive the linked score portion of the single score. This process is described in detail in technical documentation.

The nonlinear regression procedure uses a fixed set of nodes (scores on the common portion) of the tests and finds a nonlinear regression of optional scores for the common section scores. The nonlinear regression procedure appears to fit a linear relationship between nodes and appears to result in a piecewise linear function.

No statistical rationale is given for the particular choice of nodes in the regression procedure. In addition, no statistical rationale is given for the use of what is apparently a piecewise linear regression function. Alternate sets of nodes would likely lead to different linking results. In addition, a cubic spline regression procedure (i.e., smoothing splines) likely would be an improvement to the procedure used here because cubic splines (see Kolen and Brennan, 2004, for a discussion of the use of cubic splines in equating) produce a continuous curvilinear regression function and criteria exist in the literature for choosing nodes and smoothing parameters.

No statistical rationale is provided for calculating a single score by summing the scores on the common section and scores on the linked optional section. The correlation between the common portion and the optional portion will have a substantial effect on the variability of total scores. For these tests, the correlations between the common portion and the optional portions were nearly equal for the three optional sections, so the variability of total scores likely were similar. However, if at some point in time these correlations were to differ

substantially, the variability of the summed scores could be quite different for examinees taking different optional sections.

The statistical criteria used are not stated for the procedure, which gives rise to the following questions: What is the method intended to accomplish from a statistical or psychometric perspective? What statistical or psychometric assumptions are being made?

From the analyses presented, it is difficult to ascertain the extent of comparability of total scores for students taking different modules. Based on the way the procedure is implemented, it appears that the single score for a student who took a Biology optional module is considered to be comparable to a student who took a Chemistry or Physics optional module. The rationale for this comparability is not clear.

It would be informative to regress outcome variables such as college grade point average (on comparable college science courses) on the Science single score for the groups taking the Biology, Chemistry and Physics optional modules. If the total scores are comparable, these three regressions should result on regression coefficients approximately equal.

Recommendations:

94. The international PSU evaluation team considers the rationale for PSU Science score linking to fall below international standards. The process is not only incorrectly labeled but also the documentation is incomplete and the evidence of technical adequacy insufficient for high-stakes decisions. **The evaluation team recommends developing separate Science tests for Biology, Chemistry and Physics with specific purposes and intended populations in mind so that the scores would have unambiguous meaning.** Each of these tests would be reported on separate PSU scales, following standard processes already available for PSU tests. Furthermore, once the current cumbersome linking process had been replaced, the year-to-year maintenance of the new PSU Science scales through equating would be more rigorous and, hence, more defensible.
95. **Until our recommendation can be implemented, there is a need for more documentation for the current Science tests that informs the public and technical reviewers alike about the current policy decision to report a single score for Science, its rationale, and the research that informed that decision.** The practice of linking test scores from different content areas has been performed to achieve comparability through scaling. This form of linking is weak when compared to equating. For that reason, evidence of the generalization of conversion tables should be provided for sub-groups, occasions, and tests. **Other recommendations for the current PSU Science tests include the following:**
    a. **Refer to the process as "linking" rather than "equating."**
    b. **Link total scores, rather than using the current process of linking scores on optional sections and then summing the linked scores with the scores on the common portion.**
    c. **Consider using standard linking methods, such as chained equipercentile and frequency estimation equipercentile. Smoothing methods should be used with these procedures.** Thoroughly compare the results for all methods considered. Provide statistical and psychometric criteria that indicate what the procedure is intended to accomplish.
    d. With the current linking methods there is an implicit assumption that total scores are in some sense equivalent regardless of the optional module taken.

This assumption, which appears to be unrealistic, needs to be thoroughly investigated. **For any procedures considered, including these standard ones (e.g., chained equipercentile and frequency estimation equipercentile), it is important to check on score comparability.** Regress outcome variables such as college grade point average on total score for the groups taking the Biology, Chemistry, and Physics optional modules. If the total scores are comparable, these three regressions should be approximately equal.

e. **Estimate reliability, standard errors of measurement, and conditional standard errors of measurement using procedures that have been developed for composite scores. Calculate standard errors of the linked scores using bootstrap procedures.**

f. **Document processes describing quality assurance and quality checks for PSU Science score linking.**

Objective 1.3: Evaluation of IRT models for item calibration, test development and equating purposes

Description:

This objective included an evaluation of IRT methods to calibrate items and of the prospect that successive PSU forms might one day be equated.

Method:

The evaluation team read DEMRE documentation and related information about the PSU calibration process. Additional information for this objective was obtained by means of interviews with relevant stakeholders from DEMRE on March 26 of 2012, and included the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process. In preparation for the meeting, the evaluation team created a synthetic data array with 3000 simulees and 50 items with known item parameter properties utilizing standard simulation methodology.

Findings:

The evaluation team after considering all facets and elements inspected for Objective 1.3 rejects the IRT documentation and processes currently used in the PSU testing program. The documentation reviewed is misleading. The processes taking place need to be labeled properly, and processes that are not taking place need to be identified. For example, the evaluation team became aware of the misleading use of equating terminology (e.g. anchor set, model fit) and corrected their initial understanding. Specifically, the PSU testing program does not maintain a reporting scale on a yearly basis nor does it calibrate pilot items to a common scale.

A clear description of a process to calibrate piloted items is an important practice often found in mature testing programs involving high-stakes assessments. When developing and assessing test items using item response theory (IRT), practitioners design processes in a way that pilot item parameters can share a common scale with existing pilot item parameters in item banks. Producing piloted item parameter estimates within a common scale allows for an "apples-to-apples" comparisons of item parameters and related IRT

information, both among multiple pilot forms in any given year and across years of operational administrations.

In the context of Chile's university admissions process, the evaluation team recommends correcting inaccuracies present in the PSU documentation for pilot equating. The existing process for developing an anchor set, as described in official documents, is unconventional, and it does not reflect international standards. For example, the length of the anchor set is below the standard ratio between anchors set and test length. The ratio mentioned in the documentation for the PSU does not reach the minimum of 25 percent of the total test length (Kolen & Brennan, 2004). In addition, the anchor sets are constructed without guidelines on how to achieve content representation. As a result, these anchor sets fall short of completely and accurately representing total test characteristics.

There are several ways that the PSU testing program does not attain a level of analysis generally expected when equating a high-stakes assessment. The most striking irregularity was the misleading information found in the documentation of the assessment program: specifically, the information concerning year-to-year equating and pilot calibrations. The international evaluation team reiterates that these activities are not taking place, even though the documentation provided seems to indicate that they are. The evaluation team emphasizes the necessity for equating activities to be carried-out for the PSU tests.

During the review of the documentation and the interviews, the evaluation team became aware of the loose usage of equating terminology and processes. In the context of equating the PSU testing program, there is a discrepancy between what is documented and what is practiced. From the interviews, it became evident what the purpose of the so-called anchor set has been in the program. For example, the term "anchor set" is being used to refer to a group of items that are added to a pilot form with a purpose other than the calibration of pilot administrations to a common scale.

There is major drawback in the current PSU item calibration process: Year-to-year calibrations are not referenced to a common scale. The absence of such a calibration effort creates concerns about the comparability of item parameters, the associated test construction activities, and the quality of item bank information. Dimensionality analyses are also absent from PSU item calibration process. It would be commendable to add this type of analysis to check model assumptions.

The anchor items are essentially useless since they do not fulfill their purpose. Even at the design stage, the anchor set shows multiple flaws. Anchor set content and statistical characteristics are deficient in light of the international standards. The anchors are too short and under representing the total test content. The implementation of such anchor specifications, if pursued, would render bias equating results and large equating error. Also the design of the anchor sets shows lack of awareness information about the actions taken, at least from a design perspective, to reduce context effects (e.g., retaining item position of anchor items) on performance of anchor items and processes to perform checks on anchor item parameter drift. To minimize context effects, anchor items should retain their position on the tests. The lack of plans for incorporating screening of item parameter drift presents potential threats to equating accuracy.

Recommendations:

The international PSU evaluation team considers PSU equating to be below international standards. The documentation for PSU equating is not only misleading but also incomplete and inaccurate in some areas. For a national entrance exam with high stakes for thousands

of applicants, the technical adequacy is insufficient, which means that erroneous outcomes (i.e., decisions) may occur. The evaluation team proposes a refocus of efforts that would address the following recommended improvements.

96. In order to replenish the item bank as new tests are created each year, newly developed items must be field tested and equated onto the scale of the original form. Once the field test items are administered, it is necessary to place their item parameters onto the same scale as the original form of the test in order to enable pre-equating during the test assembly process. Calibration of field test item parameters can be performed with approaches reviewed by Kolen & Brennan (2004).

97. **In order to retain scale properties and allow comparability of test scores between years of test administrations, newly administered PSU tests need to be equated to compensate for differences in difficulty.** A statistical equating simply establishes the relationship between two test forms. Typically, this is accomplished through the use of a common element across test administrations—either common persons or common items. In some cases, where appropriate, an assumption may be made that two separate groups taking two separate test forms are randomly equivalent. In most university admissions test contexts—where the goal is equating test forms from year to year—a common persons design is not typically feasible. Kolen & Brennan (2004) have an extensive treatment of alternative approaches to conduct test equating that can be consulted. It is important to emphasize that test equating is not the solution to test construction issues. The test construction process aims to develop a test form that is equivalent in content and difficulty to other forms administered in previous years. Equating is a tool that compensates for differences in test difficulties that could not have been controlled during test construction.

98. **We recommend that the PSU equates test forms across test administrations.** The lack of equated scores undercuts the ability to develop assessments that are fair to test takers. Fairness could be at stake when students taking PSU test on year 1 are advantaged with respect to those who took another PSU test on a subsequent year. For an assessment to be considered fair, test scores should not depend on the particular test form taken. In Chile, PSU test scores can be utilized up to two consecutive years as part of the admission process. Equivalency of the PSU scores between forms is a necessary condition to support such an emergent use.

99. **The design of the anchor set should comply with international standards. The design should describe targets of content coverage and psychometric representation of the anchor set in such a way that the anchor set can be seen as a mini-version of the total test. The design should describe measures to control for content effects and potential drift of item anchors.**

100. The 2PL model is currently being used for item analysis in the PSU program without a rationale. **If this model is to be used in the future for item analysis or for additional purposes, the evaluation team strongly recommends following international practice and validating the adequacy of its use over the typically used alternatives of the Rasch or 3PL models.** Similarly, the evaluation team recommends developing documentation of item calibrations that are available to staff who participate in the calibration of items. Such documentations could be used to train staff on PSU item calibration processes that have been approved by DEMRE and the technical advisory committee (CTA).

Objective 1.4: Evaluation of software and processes utilized for statistical analysis and item bank

Description:

This objective included an evaluation of the software used for item analysis and item banking of the PSU. The evaluation team ascertained whether the software afforded the proper level of security as well as the faculty of providing appropriate version control of the items, the item status, and the associated psychometric data.

Method:

The evaluation team analyzed the software packages used by DEMRE for item banking and statistical analysis. Additional information for this objective was obtained by means of interviews with relevant stakeholders from DEMRE on March 26 of 2012, and included the director of DEMRE, the general coordinator, the head of the research unit and his team, as well as the head of the admissions process.

Findings:

The item bank database seems to be well designed with respect to the security of the items. The limitations placed on users minimize the possibility of security breaches for both operational items and forms. For example, only test authors can view items, authors can only view items associated with their subject areas, hardware-based keys are required for access to item images, psychometricians cannot view items or keys, and IT technicians, including database administrators, cannot view items or keys. There are options for authorized users to export the item images to *.DOC files that are saved on their local machines, so security of the tests should include ensuring the security of those authorized users' computers. These measures might include such things as encryption of those machines' hard disks, requiring that screen savers with passwords be enabled, and limiting networking of the machines to internal LANs only.

Version control of item images is present but seems weak based on the available documentation. Only authorized users can make changes, but it does not appear that these changes are tracked in any fashion or that previous versions of the items are retained. Items are locked at certain stages in the item development and usage cycle.

There is no documentation of version control of item statistics beyond an item status indicator that shows how many times an item has been used operationally. It appears that only the statistics from the most recent administration are retained in the database.

BILOG 3.11 and DIFAS pieces of software both use well-researched and recognized statistical procedures to estimate IRT item parameters and Mantel-Haenszel DIF statistics, respectively. BILOG 3.11 is limited to dichotomous items which is the item format currently in use for the PSU tests. As we have stated before, DEMRE relies on the 3.11 version of the BILOG 3.11 software. The evaluation team recommends developing an updating plan for software version updates. The BILOG 3.11 software may need to be replaced by other software, depending on future decisions. If in the future it is decided to include polytomous items on the PSU test, BILOG 3.11 would fall short when handling this item format type. Commercially available software that allows for polytomous items (e.g., MULTILOG) can be considered and evaluated. Likewise, if in the future it is decided to include IRT equating while allowing for the estimation of item parameters for more than one group, the BILOG

3.11 software should be replaced by software that allows for such kind of analyses (e.g., BILOG MG).

DIFAS accommodates multipoint items but only provides statistical measures of DIF without heuristic flags for such items. The item bank would need to be extended to provide such flags should multipoint items be added to the tests at some point in the future. In addition, the evaluation team recommends developing a plan for the software version update. (Note: More information on challenges when involving multiple DIF methodologies can be found as part of Objective 1.1.g. of this evaluation).

SAS is a robust system and is well suited for use for the Science equating analyses. The SAS code itself has sufficient documentation, and appropriate SAS PROCs are being used.

Recommendations:

The international PSU evaluation team considers the set of software tools available for analysis of the PSU university admissions testing program to be below international standards. The evaluation indicates that though there may be just enough automation *within* a functional group, there is not enough *among* functional groups. For a national entrance exam with just a single operational administration of six assessments, the processing environment may be tolerable. However, the PSU testing program would be challenged were it called upon to allow for multiple test administrations, something which occurs regularly in many university admissions programs internationally. The evaluation team proposes a refocus of efforts that would address the following recommended improvements.

101. The current systems seem to be somewhat disjointed, with much manual manipulation of item and test data required. One of the verification steps during test construction is to check the item codes to verify that they exist in the database, implying that the test author must type in the item codes manually. This is a place where an error can occur if the test author mistypes an item code and the mistyped code happens to match another (unwanted) item in the bank. The SAS code used to implement the Science test equating appears to require manual editing for each successive year. There are many references to "importing" and "exporting" data to and from the database; however, to the extent that these functions require manual manipulation, these are steps in the process where errors can occur. **It is recommended that common processes be automated as much as possible and that analyses be standardized to eliminate or reduce the amount of manual intervention required.**

102. **Versioning of both item images and item statistics can be improved.** The ID of users making modifications to items should be tracked. In addition, previous versions should be retained, both to provide a historical record of changes made to an item and also as a safeguard to allow reversion to an earlier version if needed.

103. **Statistics from all administrations of an item should be retained, and users should be able to view them together in chronological order.** Large changes in item statistics from one administration to another can indicate such things as item exposure, printing errors, or other problems. The documentation indicates that key-check analyses are performed for piloted items; key-check analyses are recommended for all items to control for unrecorded changes in items, detection of printing or production errors, errors in the importing and exporting of data to and from the database or other unforeseen errors.

104. BILOG 3.11 software can be run in batch mode, and command files can be produced automatically by custom-written software or via SAS programs, both of which are recommended. BILOG is a rather old program, and newer software might bring benefits. If using IRT for equating is possible in the future, it might be worthwhile to upgrade to newer software.

105. DIFAS does not allow for the use of command files and therefore cannot be run in batch mode. Mantel-Haenszel analyses can be easily coded in custom software or can be run in SAS. Replacing DIFAS with a solution that can be more easily automated would reduce the need for the labor involved in running DIFAS, saving the results and then importing them into the database.

106. The general process for producing the normalized scores was described in the documents available for this review, but the specific procedures and software used to accomplish these analyses were not. In general, the preceding recommendations also apply to the processes and software used for the derivation of the normalized scores. To the maximum extent possible, these processes should be automated so that they can run without manual user intervention, and software amenable to running in batch mode (such as SAS) should be used for these derivations.

Objective 1.5: Evaluation of the delivery process and of the clarity of information regarding those examined and the different users of the admissions system

Description:

This objective included an evaluation of the PSU score reporting that entailed the summary and analysis of interview responses obtained from key PSU stakeholders (students, high school teachers and university admissions officers) with respect to particular features of the PSU score reports.

Method:

The evaluation team interviewed groups of intended audiences of the reports to gather information on whether these stakeholders understood the information contained in the current reports and if they thought there was any need for improvement of the reports and processes. The interview questions were targeted to three distinct groups based on the information they currently receive from DEMRE: students, high school teachers and university admissions officers.

Findings:

The overall theme from interviews with the students with respect to their reaction to the *PSU Delivery Report Results* is that while they understood the intent and basic purpose of the report, they were not able to gather much useful information from it. The students did not believe the report contained information that would help them answer any quantitative questions about scores in the report. Many of the students wanted more specific, diagnostic feedback from the report.

With respect to the *Chilean University Admissions Report*, the students did not feel the report gave enough if any information for them to be able to answer any quantitative questions from the report. Basically, the subjects felt the report was of little help and did not provide enough information.

The Teachers' responses to the questions asking about uses and applications of the *PSU Statistical Reports* yielded a variety of answers. Some teachers felt the reports could be used to project how the students might do in the future, while others believed they could be used to measure how well the students knew the content at the time of testing. Subsequent questions asking the teachers to use the information from the *Reports* indicated a general difficulty finding the targeted information. Distributing the reports faster and more efficiently were two requests from the panel.

Overall, the university admissions officers indicated they were "satisfied" with the current process that was used and for each specific stage of the process. While subjects gave different rankings to the importance of pieces of information from the reports, they expressed a desire to get the information sooner because of tight deadlines. One also mentioned that the information they were given is much better than the information that was received previously.

Recommendations:

107. **Because students did not understand the PSU scale scores that were presented to them, we recommend that DEMRE provide additional interpretive information explaining the PSU Delivery Report Results.**

108. **Although DEMRE publishes information on weighting and admissions scores, we recommend that this information should also be tailored for inclusion in each student report.**

109. We recommend that the reports be redesigned to make it easier for students to find information, such as the number of spaces available in university departments.

110. **We recommend that the information provided regarding the areas of test takers' strengths and weakness on each PSU test in the PSU Statistical Reports for educators be suspended until the results are carefully scrutinized to ensure the reliability and validity of such information.**

111. **Because educators indicated that they use the results of the PSU tests for purposes other than university admissions, we recommend that the PSU Statistical Reports explain what the intended uses are for the PSU tests and warn against the unintended uses.**

112. Although the report contains a great deal of information, much of it quite valuable, it is difficult to find specific data to answer specific questions. For example, the report provides in the appendix the number of students admitted into university from their particular high school. However, educators were unable to locate this information. We recommend including a detailed table of contents that would improve the value of this report, making such information more readily available to educators.

Evaluation Objectives 2.1–2.4: The PSU Validity Studies

The objectives in this section are all related to the validity of the PSU tests. The way that PSU test items relate to one another—the internal structure of the test—is examined in Objective 2.1. Analyses to determine if the PSU test content fully reflects the domain it is trying to assess—content validity—is provided in Objective 2.2. Objective 2.3 considers the trajectory of PSU test scores over time. Finally, the ability of the PSU test scores to predict other important measures such as university grades and graduation rates is the subject of the study for Objective 2.4.

Objective 2.1. Internal structure of PSU exams: goodness of fit of PSU test scores analyzed with item factor analysis and item response theory models

Description:

In developing tests and intended test uses, it is often useful to develop the conceptual definitions first and then to find ways to define them operationally. Cronbach and Meehl (1955) introduced the concept of construct validity to study whether test scores are sufficient to recover components highlighted in the conceptual definitions of the tests.

Validity research on the overall dimensionality of the PSU tests and its invariance across subpopulations is scarce. There has been no study conducted to date in which item factor analysis and item response theory models were used to investigate the properties of the PSU test scores. In addition, there has been no study of invariance of a unidimensional structure between PSU forms.

The purpose of this study was to examine the internal item structure of the PSU test battery with item factor analytic and IRT frameworks, utilizing PSU data from the 2012 admissions process.

The following research questions oriented the investigation:

- What is the dimensionality of the PSU tests?
- To what extent PSU test factor structure generalize over relevant subpopulations of test takers such as:
    - Gender: Male or Female
    - Regions: North (codes 1, 2, 3, 4, 15), Central (5, 13 [Metro]), or South (6, 7, 8, 9, 10, 11, 12, 14)
    - Socio-economic status: Five quintiles of the SES variable—Quintile A defines the Lower group; Quintile B defines the Below Average group; Quintile C defines the Average group; Quintile D defines the Above Average group; and Quintile E defines the Upper group. SES was computed utilizing information from applicants' family income and parental education.
    - Curricular Branch: Scientific-Humanistic or Technical-Professional
    - Type of Financing: Private, Subsidize or Municipal?

Method:

Item factor analysis was conducted separately on all PSU tests. The purpose for the analyses was to determine whether each of the PSU tests overwhelmingly represents a single underlying factor or whether any of these tests show consistent evidence of a multi-factorial structure. The presence of a strong dimension for a PSU test contributes evidence

to support the linkage of the test items to the tested underlying latent variable (Lord & Novick, 1968). Otherwise, it shows items cannot be scaled along a single dimension, thus weakening the meaning and use of the test scores. This analysis was conducted by comparing the sizes of the first three eigenvalues estimated from item level tetrachoric correlation matrices.

Differential Test Functioning (DTF) was carried out to evaluate the invariance of factor structure across relevant subpopulations. DTF is a psychometric process to investigate the relative difference on total test scores between two groups of test takers after accounting for difference on group's levels of test performance. DTF allows test developers to identify whether the test favors one group over the other and provides lines of evidence to support (or reject) test fairness. A goal when developing tests is developing measurements that function equally well across groups. Failure in meeting this goal constitutes validity evidence against the test. To the date of this writing, DTF analyses have not been part of DEMRE psychometric processes. In the future, this type of analysis should be added to the PSU item level analyses.

The international evaluation team relied on a univariate IRT framework, after checking for unidimensional PSU solutions, for analyzing the data. The absence of DTF is indicative that PSU scores are directly comparable on subpopulations, while the presence of DIF will show that item scores are inconsistent among subpopulations. DFT is arguably more important than Differential Item Functioning (DIF) because the former speaks to impact, whereas the latter may be significant for one item, but might not have too much practical impact on test results (Templin, 2009).

Item factor analyses and differential test functioning analyses were carried out with PSU data from the 2012 admission process.

Findings:

Findings from this study supported the presence of a strong latent dimension for each PSU test. The analyses revealed a single underlying dimension for each of the PSU tests (Language and Communication, Mathematics, History and Social Sciences, Science–Common, Science-Biology, Science-Physics, and Science-Chemistry. Such a finding is encouraging and supports future use of one-dimensional item response theory models to set and maintain PSU test scores scale over years of test administration and to equate the PSU test scores. In Chile, PSU test scores from a given administration are valid for two years; thus, comparability of test scores is necessary for building a fair national admission testing program. The existence of a "single underlying dimension" for each of the tests is a necessary but insufficient condition for test validity. In our analysis the evaluation team found a single underlying dimension for each of the PSU tests. The evaluation team notes that the underlying dimensions are interpretable within a PSU test and, thus, the set of single underlying dimensions may not be the same one, even for the various Science tests.

Generally speaking, PSU tests show some evidence of differential test functioning (DTF). In these circumstances, it is reasonable to conclude that factor structure invariance of tests by subpopulation groups has been partially achieved. Particularly speaking the strongest evidence of DTF is seen for lower performing Technical-Professional students relative to higher performing Scientific-Humanistic students on the Science – Biology, Language and Communication and Mathematics tests and for much higher performing Private and somewhat higher performing Subsidized students relative to less well-performing Municipal students on the Language and Communication and Mathematics tests. These are specific cases that may warrant further consideration and review.

With respect to the results by PSU test subject, Mathematics showed the largest number of DTF flags with eight flags out of the ten sub-group comparisons. This was followed by Language and Communication with five DTF flags.

Recommendations:

113. **The evaluation team recommends adopting the IRT framework for test construction activities, item-level analyses, scaling and scale maintenance.**

114. **The evaluation team recommends selecting operational items during test construction activities so that high levels of precision at the critical decision point of the score scale are attained.**

115. **The evaluation team recommends using the factor analysis results showing the unidimensionality of the PSU to ground the use of IRT to scale and to equate the PSU.**

116. **As a result of the evaluation team's differential test functioning (DTF) analyses, it recommends that the PSU program conduct additional analyses to understand better the DTF between private and subsidized schools versus municipal schools, particularly for the Language and Communication and Mathematics tests. This is a recognized standard (Standard 7.3, AERA, APA, & NCME, 1999) for high-stakes test development worldwide where the fairness of the test across different subpopulations is an issue.**

Objective 2.2. Content validity: Logical and empirical analyses to determine if the test content fully reflects the domain and if the test has the same relevance regarding the interpretation of the scores in the population subgroups

Description:

Developing adequate content validity has always been a concern for test developers and users (AERA, APA, NCME, 1999). Typically, the most common method for ascertaining content validity has been the use of content experts in the test development process. However, the alignment of a test to a set of achievement standards is a relatively new strategy to determine the content validity of an assessment.

For the PSU to be aligned to Chile's national curriculum, it is essential that the PSU tests measures the depth and breadth of the Language and Communication, Mathematics, History and Social Sciences and Science national curriculum. Assessments that devote proportionate number of items across subsets of content and skills specified in the curriculum are better aligned relative to those focusing on peripheral content and/or disproportionate balance.

The purposes of this study were (1) to document the degree of alignment of the PSU tests to the PSU intended domain, (2) to gain deeper understanding on PSU alignment from the perspective of high school teacher and university faculty, and (3) to summarize the position of the Curriculum division of Chile's the Ministry of Education.

Regarding the first purpose, the following question guided the alignment piece of the study:

- What is the degree of alignment of PSU tests to its intended domain?

Regarding the second purpose, the following themes provided a structure for the meeting with stakeholders.

- Perceived degree of alignment of the PSU domain to classroom instruction;
- Perceived readiness of admitted entry level university students at the beginning of university instruction;

Method:

Logical and empirical analyses were carried out toward investigating degree of coverage of the intended domain of the PSU tests administered for the 2012 admissions process. This effort took into consideration Chile's national curriculum, fundamental objective (OF) and minimum obligatory content (CMO) for the Scientific-Humanistic and Technical-Professional curricular branches. The analyses were performed considering General Training and Differentiated Training (Scientific-Humanistic and Technical-Professional) utilizing Webb's (1997) alignment methodology. The method gives a comprehensive view of a test's potential to evaluate students on required material. The Webb method not only assesses the degree to which standards (CMO and OF) are addressed by test questions, but in additionally examines the level of cognitive complexity required by test questions, the breadth of knowledge required by the test questions and the evenness of coverage of standards (CMO and OF) by a test.

Pearson used the five dimensions of Webb's alignment process to judge the alignment between Chile's high school curricular content standards and the PSU tests and to answer the following question:

- **Categorical Concurrence:** Does the PSU measure what the curricular standards state students should both know and be able to do?
- **Depth of Knowledge:** Does the PSU reflect the cognitive demand and depth of the curricular standards? Is the PSU as cognitively demanding as the standards?
- **Range of Knowledge:** Does the PSU reflect the breadth of the curricular standards?
- **Balance of Representation**: Does the PSU reflect the full range of the curricular standards?
- **Source of Challenge:** Does the PSU reflect cognitive demands extraneous to those in the curricular standards?

An alignment study was performed with a Pearson team of content area specialists, all of whom are fluent in Spanish. Of the six panelists, two have traveled extensively or lived in Chile. Four of the panelists are teachers with an average of eight years working in the classroom or with curriculum development. Four of the panelists have earned advanced degrees (masters or doctoral) in their subject areas.

The Webb alignment of the PSU was performed as follows. First, the project team leader assigned the different portions of the PSU (Mathematics, Science, etc.) to the content area specialists who had been extensively trained in the Webb alignment method. The content area specialists' first task was then to assign depth-of-knowledge (DOK) ratings to the objectives of the Chilean Ministry of Education's high school content standards. They then similarly assigned DOK ratings to PSU assessment items. The panelists then determined the categorical concurrence, DOK consistency, range-of-knowledge correspondence, balance of representation, and source of challenge criteria as defined above. Once completed, the content area specialists sent their work to the project team leader, who then sent it out for independent review.

In a separate effort, university professors and high school teachers participated in a series of interviews. The evaluation team proposed using accessible groups of stakeholders identified from an initial list from MINEDUC to gather their anecdotal information on the PSU test domain and the relationship to levels of knowledge and skills relevant for entry level students to be successful. A total of 27 stakeholders participated in the interviews. Eleven university professors were interviewed. They were all from state, metropolitan universities that were part of the CRUCH. There was one teaching director, four Mathematics professors, one Physics professor, two Chemistry professors, two Biology professors, and one professor of Language and Communication. Sixteen high school teachers were interviewed. The teachers came from two private, metropolitan high schools that followed a Scientific-Humanistic curricular branch. Among the teachers were four that specialized in Language and Communication, four in Mathematics, three in History and Social Sciences and five in Science. The Science teachers consisted of two Chemistry teachers, two Physics teachers and one Biology teacher.

Interview meetings began with a high-level introduction of the purpose and ground rules followed by a general overview of the PSU evaluation. Following this presentation interviewees were asked to familiarize themselves with the OF and CMO on Chile's national high school curricular branches (Scientific-Humanistic and Technical-Professional) and to follow the directions stated in the interview protocols. The evaluation team facilitator encouraged discussion and alternate points of view from the panel members. Interviews responses were analyzed separately by stakeholder group and results categorized by major and minor findings in summary tables.

Findings:

The results of the alignment study indicate that for almost all of the PSU tests, the level of alignment of the PSU to both the Fundamental Objectives (OF) and Minimum Obligatory Contents (CMO) of the Chilean curriculum was uniformly low. One aspect to take into account when interpreting these results is that there are certain strands within the standard sets that are impossible to assess in a multiple-choice format exam of the PSU. This would tend to lower the alignment. Nevertheless, although the existence of these strands within the standards will result in artificially lower alignment scores for the PSU tests, the results of the study demonstrate that much improvement could be made.

In addition, analyses of anecdotal information from high school teachers and university professors showed that there is a fundamental disconnect between the purpose and use of the PSU to select students for university admissions and the content of the PSU that is based on Chile's high school Curricular Framework (*Marco Curricular*). This disconnect between the purpose and the use was found to be stronger for the selection of students from Technical-Professional curricular branch than for those from the Scientific-Humanistic curricular branch.

Regarding the use of the PSU as a predictor of success in higher education, a major theme from the interviews was that the PSU does not capture all of the aptitudes (i.e., abilities) needed to do well in higher education. For example, university professors stated that PSU tests do not capture student motivation or other important qualities and that students who do poorly on the PSU tests may go on to be successful at the university. Other themes that were captured included the perception that students with lower socio-economic status (SES) were at a disadvantage on the PSU tests and that teachers in 11^{th} and 12^{th} grades focus more on teaching to the PSU tests than to teaching the curriculum.

Finally, the Curriculum Unit of MINEDUC provided their own analyses outside of the evaluation work on the National Curriculum of Chile and its relation to the PSU.

- Consistent Alignment: the main concern, which derives from the report requested by DEMRE itself, is that the PSU evaluation framework uses as reference one section of the curriculum, the CMO, which do not necessarily render account of the total extent of the curriculum. In other words, the PSU utilizes a methodology which aligns it only superficially with the curriculum, leaving aside that which is central and the fundamental aspects of the curriculum.

- The Chilean Curriculum, as any modern curriculum, has recently been updated and revised. Specifically, the Secondary Educational curriculum was modified significantly in 2009: This change has been implemented gradually year by year. Therefore the existence of certainty and transparency is fundamental, for the whole educational system, in relation to which curriculum is being evaluated and how the intersections between two curricula are constructed.

- To date, 45% of secondary school enrollment corresponds to the Technical-Professional curricular branch. An increasing number of graduates of this curricular branch takes the PSU as part of the admissions process into higher education. From Curriculum Unit's perspective, there is a concern about the distance between that which is declared (the PSU as a general assessment) and that which is real (the PSU as a general and differentiated assessment, which emphasizes the Scientific-Humanistic curricular branch). [MINEDUC, personal communication, January 2013]

Recommendations:

117. **We recommend a review of the policy of using the Curricular Framework as the basis for the development of the PSU test frameworks.** As a part of this review, we recommend the development of a framework that describes the aptitudes (e.g., abilities) and relevant non-cognitive variables (e.g., study skills and motivation) needed by students in order to be successful at the university. Such a framework would focus the PSU on the aptitudes necessary to succeed at the university and complement the measure of high school achievement found in NEM and combined together in the postulation score.

118. Although the evaluation team has recommended aligning the PSU tests to standards for success at the university, we acknowledge the urgency to develop the 2013-14 test forms based on the full implementation of the 2009 curricular reform. To that end, the evaluation team recommends performing an alignment study on these PSU test forms. The results of this study should inform the broader recommendation to redirect the emphasis of the PSU to university success.

119. We recommend reviewing the item types used on the PSU tests to address the perceived low level of cognitive complexity found on the tests due to the exclusive use targeting of low-order thinking-skills and minimum obligatory contents.

Objective 2.3. Analysis of trajectories of PSU scores for subpopulations throughout time, considering dependence, mode and gender

Description:

Because admission to colleges and universities decisions involves the use of admission test data, analysis of the trajectory of college admission test scores through time becomes an important element in the portfolio of institutional validity studies to support generalizations

across time. Analysis of trajectories of admission test scores contributes to develop better understanding of subgroups' test performance across years of test administrations and to spot suspicious downward and upward trends. When disaggregated by specific subpopulations, the analysis of test scores trends becomes a powerful tool to monitor performance on college admission tests and to inform policy decisions in order to close observed gaps in the test performance of various subpopulations.

The primary purposes of the research were to analyze trajectories of PSU scores over time and to pinpoint variables that moderated those trajectories. An analysis of the trajectory of university admissions test scores over time is an important element of institutional validity studies because such longitudinal analyses aid in the identification of stable trends on test performance and in pinpointing gaps on test performance for relevant subpopulations.

The evaluation team researched the following questions:

- What is the trend in PSU tests scores for the following subpopulations?
    - Gender: Male or Female
    - Region: North (codes 1, 2, 3, 4, 15), Central (5, 13 [RM]) or South (6, 7, 8, 9, 10, 11, 12, 14)
    - Socio-economic status: Five quintiles of the SES variable—Quintile A defines the Lower group; Quintile B defines the Below Average group; Quintile C defines the Average group; Quintile D defines the Above Average group; or Quintile E defines the High Upper group. SES was computed utilizing information from applicants' family income and parental education.
    - Curricular Branch: Scientific-Humanistic or Technical-Professional
    - Type of high school: Private, Subsidized or Municipal
- What school-level variables moderate the relationship between PSU scores and NEM?

Method:

The study relied on longitudinal data sets spanning the admissions process from 2004 through 2011. DEMRE provided databases with applicants' demographical information, PSU test scores and high school grade point average (NEM). Descriptive information such as n-counts, mean, and standard deviation were computed and reported for PSU scale scores (and raw scores) by test and admission year for the total population of applicants and for each of the subpopulations. As part of the descriptive statistics, plots of mean scale scores (with 95% confidence bands around mean scale scores) were generated and interpreted. Additionally, hypothesis testing analyses were carried out with analyses of variance with one dependent variable to test for null differences on PSU mean scale scores by year and by subpopulation. Analyses of variance results were summarized in tables and interpreted.

For the second research question, the effects of school-level characteristics on levels of covariation between PSU subtest scores and NEM were studied with hierarchical linear modeling (HLM) (Kreft & De Leeuw, 1998; Raundenbush & Bryk, 2002). The model allowed for an investigation of the extent to which school-level characteristics (i.e. type, curricular branch) affected the relationship between NEM and PSU scores. The following variables were in the model.

- Criterion: PSU scores
- Level 1 unit of analysis: Student
    - Student-level predictor: NEM

- Level 2 unit of analysis: High School
    - School-level predictors: School-level SES, Curricular Branch, School Type, Region, and % of student population that is Female.

Findings:

Results of the trend analysis indicated that, on average, PSU scores remained fairly consistent over time with a slight upward trend beginning in 2007. An examination of subpopulations indicated that this upward trend is largely due to the performance of Private schools and schools with a Scientific-Humanistic curricular branch. Trend lines disaggregated by school type and curricular branch showed that scores steadily increased over time for Private schools and schools with a Scientific-Humanistic curricular branch, while the scores stayed flat for the Municipal school type and for school with a Technical-Professional curricular branch. In addition to differences in scores due to school type and curricular branch, the gender, socioeconomic status (SES) and the region in which the student resided significantly moderated the trend of PSU scores. The patterns of PSU test scores across relevant demographic variables are akin to patterns observed internationally with university-bound seniors. The gap size, on the other hand, is large for the Chilean population of university bound students, particularly for the subpopulations based on type of school and socio-economic status.

A secondary focus of this research was to examine the covariation between high school performance grade point average and PSU scores. It was discovered that, while NEM predicted performance on all PSU subtests, it performed particularly well for Mathematics and Science. School-level variables that moderate this relationship between NEM and PSU test scores were also examined. School type and curricular branch were particularly strong moderators of this relationship as the slope for NEM was substantially steeper for Private schools and schools providing the Scientific-Humanistic curricular branch. SES, region, and the percentage of Females at a high school also moderated the relationship between NEM and PSU scores.

Recommendations:

120. **The evaluation team recommends carrying out test score equating on a yearly basis.** Along this line, we recommend careful inspection of comparability PSU test scores between years. In Chile PSU test scores can be used for two consecutive admission processes. After accounting for lack of test score equating and changes on PSU test specifications (e.g., mathematics test increased length in the 2012 admission process), equity of PSU test score between adjacent years is at stake. It should be a matter of indifference for applicants whether they take the PSU test in 2011 or in 2012.

121. **The evaluation team recommends inspecting invariance of equating functions across relevant subpopulations of applicants.** These types of verifications are germane to developing validity evidence on meaning of test scores. Strong equating results should be invariant across subpopulations; otherwise, linking studies to align score scales should be performed to allow for comparisons of PSU test scores.

Objective 2.4. PSU predictive validity: To complement predictive validity on population groups throughout administration years, considering the differences experienced in those taking the PSU and the test variations since its implementation (2004), which shall contemplate a differential validity analysis

and possible differential prediction of the PSU through year and type of career, considering subgroups defined by gender, dependence and education mode

Description:

Predictive validity refers to the ability of test scores to forecast performance on a relevant criterion (AERA, APA, NCME, 1999). University admission decisions are complex and involve multiple measures among which university admission tests scores and high school academic performance are often the focal variables (i.e., predictors) while university academic performance measures are the outcome variable of interest (i.e., criterion). Examples of university academic performance are first semester grade point average, first year university grade point average, and cumulative university grade point average. Internationally, a long lasting recommendation for general admission purposes considers university admissions tests and prior academic record (e.g., high school grades) to be useful predictors of university grades. The best prediction models often involve all of the above predictors.

The purpose of this study was three-fold: (1) document the ability of PSU test scores and high school academic performance (NEM and high school ranking) to predict university students' academic outcomes; (2) document incremental prediction value of the variable ranking; and (3) examine the extent to which PSU test scores and high school academic performance exhibits differential prediction for relevant demographic variables. The study documented PSU predictive validity performance for overall (all careers) and career-type levels.

For predictive validity, the study sought to answer the following research question:

- What is the predictive validity of PSU test scores and high school grade point average (NEM) on university first-year grade point average, second-year grade point average and university graduation?

Incremental predictive validity is the degree to which a variable better predicts outcome than an alternative variable. The incremental predictive validity analysis sought to answer the following question:

- What is the incremental predictive validity of the variable ranking (measured as a proxy variable from NEM) over and beyond PSU test scores and NEM on university first-year grade point average, second-year grade point average, and university graduation?

Differential predictive validity is another kind of institutional validity study that is directed to investigate the degree of similarity and difference in predicted outcomes among relevant subpopulations. With respect to differential predictive validity, this study sought to answer the following question:

- What is the differential predictive validity of PSU test scores and high school grade point average on university first-year grade point average, second-year grade point average, and university graduation for the following variables:
    - Gender,
    - Socio-economic level,
    - Region,
    - High school curricular branch, and

- Type of high school funding?

Method:

The investigation made use of longitudinal data sets for university admissions that spanned 2004-2012. DEMRE provided the databases with PSU test scores and high school grade point average (NEM), and MINEDUC provided the databases with students' university academic outcomes and their high-school ranking score. MINEDUC also provided a list with classification of university careers by career type used in the research.

Linear and logistic regression analyses were run separately for each career within a university and summarized across careers and universities. Corrections for restriction of range, involving variances and standard deviations of PSU test scores from the population of university-bound seniors (i.e., population of university applicants), were applied to the Pearson validity coefficients from the population of university students. PSU predictive validity results were weighted so as to assign more weight to larger sample sizes. Incremental prediction validity of the variable ranking was computed by fitting base and revised models to the data set. The revised model used ranking as an additional predictor. The effect of the variable ranking was documented by computing the difference in variance reduction of university outcomes (e.g., the revised model minus the based model). Analyses of differential validity were carried out by demographic variables with estimates of standardized residuals computed within careers and disaggregated by demographic variables. For summative purposes the individual students' residuals were averaged across careers and admission years before their disaggregation by demographic variables.

Findings:

The findings for the prediction study indicate that PSU tests have to a certain extent, the ability to predict university outcomes, particularly for first- and second-year grade point average. However, the prediction values found were smaller than those reported internationally. The variable "rank in high school" contributed to the reduction of the uncertainty of predicting university outcomes after controlling for PSU test scores and NEM. The largest amount of variance reduction happened for university completion. All in all, the PSU test scores and high school performance measures appear to result in comparable amounts of differential prediction validity for major demographic variables.

Career-type level results showed similar trends observed in the overall analyses. When examining predictive and incremental validity results by type of career, we saw similar prediction patterns to those from the overall analyses. For example, PSU Mathematics and Science scores and high school academic performance (NEM and ranking) showed larger predictive capacity than PSU Language and Communication and History and Social Sciences scores. In addition, applicants' high school ranking showed incremental predictive validity of university outcomes (over and beyond PSU test scores and NEM), although its contribution was smaller than the one found from the overall analyses.

Recommendations:

122. **The evaluation team recommends continuing developing lines of supporting evidence for the use and meaning of PSU measures.** Several new predictor measures should be added in future revisions of Chile's admission criteria after their careful evaluation to reduce amount of uncertainty when predicting university outcomes. **In this context, we recommend conducting validity studies to establish lines of evidence to support decision-making process to move forward with intended changes.**

123. We recommend investigating alternative criteria for predictive validity research beyond first-year university grade point average or graduation rates by including measures of continuing studies in graduate school, of being hired in career-related occupations and of entry-level salary.

124. We recommend investigating whether the university grading practices are uniform within a career at universities and across the same career at different universities, as this information would further ground the findings of predictive validity measures that use university grade point average as a dependent variable.

PSU Test Evaluation by Validity Question and General Recommendations

In this section, the evaluation team summarizes, through a series of questions, the validity of the PSU test battery. The discussion is not a summary of the comprehensive evaluation results. Rather, the questions address in this section focus on very important aspects of the PSU: the content of the test, its reliability, the constructs measured by the test and the uses and interpretations of the test. These questions are intended to provide an international frame of reference to highlight areas where the test should be improved.

Question #1: Do the PSU tests strongly align to the intended domain?

This question was in part investigated through an alignment study that focused on the degree of alignment of PSU tests to their intended domains. In addition, interviews with PSU stakeholders (university professors and high school teachers) were conducted to determine the perceived degree of alignment of each PSU domain to classroom instruction. The findings of the study and interviews, noted in Objective 2.2, indicate that none of the PSU tests was strongly aligned with its intended domain.

Question #2: Does the difficulty of the PSU tests adequately target the applicants' level of ability?

This question was investigated with a demonstration performed by the evaluation team in Objective 1.1.i., which examined the latent ability distributions of the PSU tests. The evaluation team found that the only test that adequately targeted the applicants' level of ability was the Language and Communication test.

Question #3: Do the PSU tests accurately measure applicants' ability at the most important regions of ability?

For this question, the evaluation team analyzed in Objective 1.1.i the conditional standard error of measurement (CSEM) for each PSU test. The results of these demonstration showed that maximum information (i.e., low CSEM) would be obtained at a high (i.e., selective) level on the ability scale for each test at which admissions decisions would likely occur. Science and Mathematics were found to meet this low CSEM criterion within the range that is relatively selective. However, the CSEM increases dramatically outside of this range (e.g., at the very most selective range of ability, even for Science and Mathematics).

Question #4: Does each of the PSU tests reflect an underlying unidimensional trait?

The findings of the item factor analysis conducted separately on all PSU tests supported the claim that there is a strong latent dimension for each PSU test. The analyses in Objective 2.1 revealed a single underlying dimension for each of the PSU tests (Language and Communication, Mathematics, History and Social Sciences, Science–Common, Science-Biology, Science-Physics, and Science-Chemistry).

Question #5: Do the PSU items and tests perform the same way across major subpopulations?

The differential item functioning (DIF) analyses found in Objective 1.1.g and the differential test functioning (DTF) analyses found in Objective 2.1 were used to investigate this question for the following subpopulations: Gender, Socio-economic status (SES), Region (Metropolitan, North and South), modality (Public, Private and Subsidized), and curricular branch (Scientific-Humanistic and Technical-Professional). Generally speaking, although

most PSU items show negligible DIF, the PSU tests show partial evidence of DTF for modality and curricular branch for all subjects except the Social Sciences test.

Question #6: Do the PSU tests strongly predict applicants' university outcomes?

This question was examined in a study described in Objective 2.4. Though the PSU Mathematics and Science tests showed medium values of predictive validity, in no instance did any of the PSU tests achieve a prediction validity index close to the lower bound of prediction validity indices observed internationally.

Question #7: Do the PSU scale and relevant cut points remain constant over years?

International standards require that test forms be equated in order to allow comparisons of scale scores across years. In Objective 1.3, the evaluation team confirmed that test score equating of year-to-year administrations had not been carried out throughout life of the PSU testing program. This finding implies that the PSU scale and cut points *do not* remain constant across years.

Question #8: Are the PSU score reports useful and clear for the intended audiences?

Interviews with PSU stakeholders (university professors, high school teachers and students) were undertaken in Objective 1.5 to explore their opinions concerning the usefulness of the score reports that they receive for the PSU. These stakeholders noted a lack of clarity and usefulness of the PSU score reports that they examined.

Question #9: Should PSU test scores be used to assign scholarships?

For this question, the evaluation team considered findings described in Objective 1.1.h and Objective 1.3. These objectives show that there are significant problems with the PSU scale scores and the cuts for assigning scholarships. At this point in time, PSU test scores are used to grant social benefits based on cut scores that need further validation.

General Recommendations

These judgments portray a testing program that is still developing and in need of improvement in several areas. While a more nuanced picture of this program can best be found by reviewing the entire Evaluation Report or by reviewing the major recommendations presented by objective in this Executive Summary, the following two general recommendations can be used to provide a blueprint for the improvement of the PSU.

General Recommendation 1: The basis for developing the PSU tests should be moved from classical test theory (CTT) to item response theory (IRT).

Item response theory (IRT) is a psychometric theory that defines and models the performance of test-takers in terms of an underlying theoretical trait or ability. On a mathematics test, for example, this would mean that the performance of test-takers on specific test items and on the test as a whole would be defined by an underlying trait called "mathematical ability". What makes IRT so powerful is that the level of the unobservable trait or ability can be inferred and estimated based on the performance of an observable set of test items. Using item response theory as the foundation for the development of the PSU would address a number of the testing program's shortcomings as presented in the table above and in the body of the evaluation report.

First of all, IRT explicitly models the relationship between a test-taker's ability and the difficulty of a given item. This is a tremendous advantage for test construction since this allows the test developer to target explicitly the difficulty of a test to the ability of a group of test takers. This alleviates the problem of creating a test that may be too easy or too hard for a target population of test takers, which is the concern raised by Question #2 in the table above with respect to the PSU tests in Mathematics, History and Social Sciences and Science. In addition, given a large set of items, the use of IRT test characteristic curves (TCCs) allows test developers to match to a high degree the statistical properties of a new test form with that of a prior test form.

Second, IRT methods provide for an examination of test information function (TIF) curves and conditional standard error of measurement (CSEM) curves. These curves show degree of measurement precision as a function of test-taker ability. This means that test developers could determine what part of the PSU score scale is showing the most accurate measurement. Through an iterative process of item selection and replacement and examination of these curves, the test developer could focus the greatest level of measurement precision at the place on the test scale where the most important decisions would need to be made (Question #3). In the case of the PSU tests, this point would be around the cut points on the PSU score scale that are used for university selection decisions.

A third important point is that using IRT would provide a solid framework for calibrating pilot items and placing them on the same scale. IRT item difficulty and discrimination parameters provide an alternative to the use of CTT *p*-values and point biserial or biserial coefficients and are sample independent. This means that when items are properly calibrated, the values of the IRT item statistics do not depend on the vagaries of the specific sample that was used to calibrate them. This is not the case with the CTT item statistics that are used for the PSU tests today.

A fourth use of IRT would allow the PSU to eliminate the correction for guessing that is currently being used. The 3-parameter logistic (3PL) model is a very flexible IRT model that explicitly takes into account item difficulty, item discrimination and a correction for guessing. By modeling guessing within the IRT framework, test developers could better understand how this affects an applicant's probability of answering an item correctly.

The PSU testing program is using IRT analyses now, albeit in a perfunctory way. The move to using IRT as the framework for the entire program could be phased in over a few years. This transition of the PSU tests from a CTT- to an IRT-based program can be accelerated if the proper technical, policy and administration resources and buy-in are in place.

General Recommendation 2: The PSU tests should be equated using IRT.

Every year, the PSU tests are administrated in new test forms. While this is essential from a test security point of view, this can give rise to another issue. According to Kolen and Brennan (2004)

> The use of different test forms on different test dates leads to another concern: the forms might differ somewhat in difficulty. *Equating* is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content. (Kolen & Brennan, 2004, p. 2)

One approach to carrying out equating is to base the statistical adjustment needed on the performance of a set of items in common across the test forms from the two administrations. These sets of common items are referred to as *anchor sets*.

Currently, the PSU program does not equate its tests which mean that the scores from those tests are not comparable from year-to-year. Although the PSU program employs what they call anchor sets that appear across test forms, they are not used do any equating of test forms.

Another problem is that the PSU program tries to equate the Science tests arising from different combinations of the common section with different optional sections (Biology, Physics and Chemistry). Because these optional sections bring content differences, scores for students taking the different optional sections cannot be considered to be equated

A solution to these problems is to use IRT, which provides a well-developed framework for equating test forms using anchor sets. This would help the PSU program in several ways.

First, using IRT item anchor sets that actually function as intended would allow for different pilot or operational test forms to be equated. This would allow for direct comparisons of IRT item statistics on the pilot and operational forms and greatly aid in the construction of new test forms.

Second, the use of anchor item sets across operational test administrations and IRT would allow the PSU tests to be truly equated from year to year. This would allow comparisons of applicant ability across administrations and ensure that the PSU scale score cut points for selection decisions on an administration correspond to the cut scores from earlier administrations.

Finally, creating separate PSU tests in Biology, Physics and Chemistry would recognize the fact that content is quite distinct and that the optional sections cannot be equated to one another. Proceeding forward with separate tests would allow for better measurement of the underlying constructs and the true equating of tests across administrations.

The move to equating the PSU test forms using IRT could be phased-in along with the other uses of IRT described above.

General Recommendation 3: The PSU program should develop an on-going research program to validate the uses and interpretations of the tests.

The evaluation team took validity as a cornerstone of the PSU evaluation. According to the current edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), validity "refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of test scores" (1999, p. 9). Validation thus refers to a process to develop evidence to support intended interpretations and uses of test scores (Cronbach, 1984; Kane, 2006; Vernon, 1963). The evaluation team performed a number of studies related to the validity of the PSU tests, and these studies can be used as models for development of on-going research program to examine key validation issues as shown above.

Question #1 asks if the PSU tests align to the domain they intend to measure. To evaluate academic competency to attend university with some measure of success, we can draw reasonable levels of inference from applicants' performance on admission tests developed to measure content standards defining that academic competency. If an adequate sample of a

test's items has been evaluated in some way and measurement error is within tolerance levels, then test scores can be reasonably accepted as measures of a level of competency attained on the intended test domain; otherwise, the degree of evidentiary support would not be as strong. The methodology in the evaluation team's study for Objective 2.2 could be used for the purpose of gathering this content-related validity evidence. This evidence will be especially important if the content-basis for the PSU tests change going forward.

Question #2 focuses on the unidimensionality of the PSU tests and was examined by the evaluation team in Objective 2.1. As described above, the evidence supported the claim of unidimensionality for each the PSU tests. Nevertheless, the PSU program will need to continue studying this issue, especially if wants to move to an IRT-based framework for test development, as unidimensionality is a necessary condition for using IRT models.

Predictive validity evidence is at the heart of Question #6. In fact, the answer to this question could in many respects be the most important source of evidence for validating the PSU tests: If the PSU tests do not predict university outcomes, then why are we using them? Currently, the predictive validity coefficients of the PSU tests are low with respect to those seen internationally. This indicates the need to continue exploring in depth the relationship between the PSU test scores and the variables associated with university success such as first-year grade point average and graduation rates.

Question #8 focuses on the quality of score reports, which is directly related to test score interpretation, while Question #9 looks directly at PSU test use. In each case, evidence for the consequences associated with the PSU tests needs to be gathered. This is especially crucial if the PSU is to be used for purposes outside of its original purpose.

In any event, the process of test validation is an ongoing one. The *Standards* state, "As validation proceeds, and new evidence about the meaning of test's scores becomes available, revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test" (AERA, APA, & NCME, 1999, p. 9). The program should formally put in place a process by which PSU tests can be validated on a continuing basis.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing.* Washington, D.C.: Author.

Comité Técnico Asesor. (2013). *Objectivo*. Retrieved from http://www.cta-psu.cl/objectivo.asp.

Consejo Directivo. (2010). *Consejo directivo para las pruebas de selección y actividades de admisión.* Retrieved from http://www.consejoderectores.cl/site/GobTrans/activa/documentos/ConsejoDirectivoPruebas.pdf

Cronbach, L. J. (1984). *Essentials of psychological testing*. New York, Harper and Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

DEMRE. (2010a). *Prueba de Selección Universitaria (PSU): Antecedentes y especificaciones técnicas.* Santiago: Universidad de Chile.

DEMRE (2010b). *Studio de Confiabilidad de las pruebas de selección universitaria*. Admisión del 2010. Santiago Chile: Autor.

DEMRE (2012). *DEMRE. Departamento de evaluación, medición y registro educacional.* Retrieved from http://www.demre.cl/demre.htm

Educational Testing Service. (2005). *Evaluación externa de las pruebas de selección universitaria (PSU)*. Princeton, NJ: ETS Global Institute.

Feldt, L., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd edition, pp. 105-146). New York: American Council on Education and Macmillan.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Norwell MA: Kluwer Academic Press.

Hofstee, W. K .B. (1983). The case for compromise in educational selection and grading. In S. B. Andersen & J. S. Helmick (Eds.) *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass.

International Test Commission. (2012). *ITC guidelines for quality control in scoring, test analysis, and reporting test scores*. ITC: Author.

JUNAEB. (2012). *Beca JUNAEB para la PSU*. Retrieved from http://www.junaeb.cl/prontus_junaeb/site/artic/20100114/pags/20100114174738.html)

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4th ed., pp. 17-64). Westport: American Council on Education and Praeger Publishers.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

MINEDUC. (2011). *Aprueba bases administrativas, bases técnicas y anexos de licitación pública, sobre servicio de evaluación de la Prueba de Selección Universitaria (PSU)* (ID Nº 592-44-LP11). Santiago, Chile: Autor

MINEDUC. (2012). *Educación superior. Aporte Fiscal Indirecto*. Retrieved from http://www.superior.mineduc.cl/index2.php?id_portal=38&id_seccion=3063&id_contenido=

Organisation for Economic Co-Operation and Development. (2009). *PISA 2006 technical report.* OECD Publishing. Retrieved from: http://www.oecd.org/pisa/pisaproducts/pisa2006/42025182.pdf

Organisation for Economic Co-Operation and Development. (2012). *PISA 2009 technical report.* OECD Publishing. Retrieved from: http://www.oecd.org/pisa/pisaproducts/pisa2009/50036771.pdf

Templin, J. (2007). *Introduction to differential item functioning*. [PowerPoint]. A presentation to the American Board of Internal Medicine for an Item Response Theory Course. http://jtemplin.coe.uga.edu/files/irt/irt07abim/irt07abim_lecture10.pdf

Vernon, P. (1963). *Personality assessment*. London, Methuen.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Objective 1.1.a. Quality, security and confidentiality standards regarding the development of items and tests: training of drafters, item drafting, revision, test assembly, printing, distribution and application

The evaluation team developed and performed interviews with relevant stakeholders from DEMRE on March 19 and 20 of 2012. The interview process took a total of 8 hours broken down into four sessions. The purpose of the interviews was to gain deeper understanding on the:

- Process to develop PSU frameworks and specifications guidelines for developing and writing items (Facet 1)
- Process to select and train PSU item writers (Facet 2)
- Process to commission writing of PSU items (Facet 3)
- Process to review and accept draft PSU items (Facet 4)
- Item authoring tools and item bank (Facet 5)
- Process to distribute and administered PSU tests (Facet 6)
- PSU Scanning and scoring processes (Facet 7)

All the interviews were performed within DEMRE offices following an agreed-upon schedule for the visit. The interviews covered the seven facets defining the evaluative objective (1.1.a) and their elements that were agreed upon with the TC during the goals-clarification meeting in Santiago, Chile, in January 2012.

The following DEMRE staff participated in the interviews:

- Head of the department of test construction
- Coordinator of test construction committees
    - Mathematics
    - Language
    - History and Social Studies
    - Science

- Head of research unit and his team
- Head of logistics unit
- Head of the educational records
- Head of computer unit
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees.

The following subsections contain the results of the evaluation for Objective 1.1.a., Facets 1-7.

The international evaluation team relied on professional standards for appraising the merits of the quality and security of the PSU process. A framework for evaluating PSU approaches for quality and security is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.1

> Test and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development. (p. 43)

Standard 3.2

> The purpose of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose of the test and about the relation of items to the dimensions of the domain they are intended to represent. (p. 43)

Standard 3.3

> The test specifications should be documented, along with their rationale and the process by which they are developed. The test specifications should define the content of the test, the proposed number of items, the items formats, the desired psychometric properties of the items and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information. (p. 43)

Standard 3.4

> The procedures used to interpret test scores, and, when appropriate, the normative or standardization samples or the criterion used should be documented. (p. 43)

Standard 3.5

> When appropriate, relevant experts external to the testing program should review the test specifications. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented. (pp. 43-44)

Objective 1.1.a. Facet 1. Framework and specifications for item development

GENERAL PROCESS DESCRIPTION

Chilean universities forming part of the Council of Rectors of Chilean Universities (CRUCH), and the eight private incorporated universities, use the University Selection Test (PSU) as the assessment tool to select applicants for openings offered by those universities. The sole objective of the PSU "consists in the elaboration of an application ranking" (DEMRE, 2010b, pg. 6), starting from the evaluation of curricular contents and intellectual skills extracted from the national curriculum, as elaborated by the Ministry of Education in 1998 regarding the four high school years.

In 2000, the CRUCH decided to develop the PSU in response to dissatisfaction with the educational reform of the 1990s, as exemplified by the Scholastic Aptitude (PAA) and Specific Knowledge (PCE) tests, which had been used for entry into the universities. It was during this period (2001–2003) that a controversy also arose about the creation of a Higher Education Admissions System (SIES) (DEMRE, 2010b).

The PSU consists of tests in Language and Communication, Mathematics, History and Social Sciences, and Science, the last of which includes Biology, Physics and Chemistry. These tests match the training areas defined by the General Training Plan established at a curricular level (DEMRE, 2010b). The PSU assesses the Fundamental Objectives (OF) and Minimum Obligatory Contents (CMO) of high school education, which are readily measurable by means of paper-and-pencil tests. The inclusion of these OFs and CMOs in the test happened progressively, from 2003 through 2006, in accordance with a plan designed by the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior,* a committee that was established in 2002 "as a relational mechanism between DEMRE, the Ministry of Education and the Technical Advisory Committee of the Governing Council" (DEMRE, 2010b, p.30).

The PSU was first implemented in 2003, with regard to the 2004 university admissions process, and its design was the responsibility of the Assessment, Measurement and Educational Registration Department (DEMRE), which acts under the auspices of the Vice-Rectory for Academic Affairs of the Universidad de Chile. Within the different operational units forming part of DEMRE, the Test Construction Unit (UCP) is responsible for the creation and management of the PSU, a duty which it carries out through the approval of the Technical Advisory Committee (CTA) of the Consejo de Rectores, an organization created by order of the CRUCH to be a "comptroller organization responsible for assuring the quality of the measurement instruments and for the transparency of the Admissions Processes" (DEMRE, 2010b, pg. 14). Among its responsibilities, the CTA is in charge of operating as a mediation organization between the Directing Council and the technical teams responsible for the elaboration and application of the admissions tests (PSU) (extracted from web page: http://www.portaldemre.demre.cl/sitios.htm).

For the purpose of the design and development of the PSU, the UCP is organized into committees, each responsible for one of the tests. So, there is a Language and Communication committee, a Mathematics committee, a History and Social Sciences committee and a Science committee. the Science committee includes subcommittees for Biology, Physics and Chemistry. Each committee proposes the design of the theoretical framework and of the technical specifications of the test assigned to it, which still must be approved by the CTA.

Given the need to ensure the articulation between the curricular contents and the assessment frameworks of the PSU, the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior* "established the principle of curricular reference" (DEMRE, 2010b). Curricular reference is understood as the inclusion of cognitive reasoning skills (at the same level of importance), which are derived from the OF, and from the CMOs, which are indicated in the curriculum for all sectors and subsectors evaluated by the test.

In accordance with the goals proposed by the DEMRE document (DEMRE, 2010b, p. 18), the notion of curricular reference refers to the assessment of two dimensions: cognitive skills assessment and minimum obligatory contents.

Apart from the concept of curricular reference, applicant profile corresponds to the description of maximum competencies to which the student may aspire once the high school education is finalized (DEMRE, 2010b, p. 19). The curricular matrix or test specifications consist of one double entry table, out of which arise spaces or boxes where the number of items to construct is defined. These specifications orient the construction of questions for each test and translate the competencies into performance indicators.

Therefore, the elements constituting these tables or matrices are, on the one hand, the specific Thematic Axes, which are derived from the CMOs and, on the other, the Cognitive Reasoning Skills, which arise from the pedagogical actions derived from the curriculum OFs. It is worth pointing out that the theoretical frameworks that currently orient the tests achieved their definitive form between 2006 and 2008 by reason of the progressive inclusion of contents prescribed by the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior*.

There exists a reality which complicates the objective of the PSU achieving complete coverage of the national curriculum; it is that the national curriculum has two branches: the Scientific-Humanistic curricular branch and the Technical-Professional curricular branch. Even though both of these curricular branches have in common a considerable amount of content, they are different from each other, especially between the third and fourth years of high school education. Since the PSU is mainly focused upon the assessment of the Scientific-Humanistic curricular branch, the students who have been trained under the Technical-Professional curricular branch may have some type of disadvantage in obtaining the higher range of scores, those which are required by the universities for the selection process.

On the other hand, the PSU focuses on the part of the curriculum called general training, rather than on the part of the curriculum called differential training, taking into account that the latter receives greater coverage in the private educational institutions than in public institutions. This focus on general training is meant to counteract a possible advantage held by private education high school graduates. Nevertheless, DEMRE, which is a technical unit and not one of research, does not study the possible differential effects due to the type of curriculum or type of institution. If such studies were performed, there might be room to consider additional criteria regarding the adjustment of the theoretical frameworks of the tests.

In summary, the PSU tests have large common frameworks, most of which correspond to the current national curriculum for high school education, though the tests somewhat favor the material found in the Scientific-Humanistic curricular branch. The national curriculum is applied over all of the national territory and, at least theoretically, the PSU focuses its assessment over the whole of this curriculum. Additionally, regarding each test forming part

of the PSU battery, the respective DEMRE UCP committees have each created a document called the theoretical framework (of the respective test), which specifies those CMOs that are measurable, attending to the nature of the instrument, that is, a paper-and-pencil test, with selection questions. That is, for the construction of the more complete version of the theoretical frameworks of the tests (a process which ended between 2006 and 2008), the UCP Committees responsible for the design of the tests had to take into account the concepts of curricular reference and curricular matrix and, under these guidelines, apply that prescribed in the plan for inclusion of contents defined by the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior*. Simultaneously, the committees had to analyze the practical possibilities for the assessment of the curriculum CMOs as well as the relevance of these CMOs regarding some tests that seek to be a nexus between the high school education and higher education in order to proceed with the selection of the OFs and CMOs, which would effectively be objects of assessment.

According to the reference frameworks of the PSU tests, the test specifications are an attempt to capture important content areas of the national curriculum. The theoretical framework of the Biology test indicates the following:

> [T]he assessment of the Biology skills and contents takes place through a number of multiple option items which considers contents and skills that are representative of the Biology subsector, but at the same time considers those contents which shall have greater relevance with respect to undergraduate scientific university careers. (DEMRE, 2011)

The reference frameworks also state that some of the curricular subjects are not included in the test specifications. As a result, the specifications of tables cannot be seen as faithful representations of the national curriculum.

> For this reason, even though the Science Test – Biology corresponds to the High School Education Curriculum, the generation of the specifications tables is not its faithful representation. (DEMRE, 2011)

At the same time, with respect to the theoretical framework for the Physics test (DEMRE, 2011, p.17), it is indicated that "for each one of the modules in which the Physics Science test is structured, the relation between OFs and CMOs is established for each one of the subject areas defined" alluding to those areas established for the different levels of high school education in the national curriculum.

Included in page 5 of the theoretical framework for the Chemistry test, it is explained that in the design of said framework, a curricular analysis was advanced which had three aspects in consideration: "Sequence: Development line for the OFs and of the CMOs throughout the four levels of High School Education; Relation between the OFs and the CMOs for each level of High School Education, and relation between the OFs, the CMOs and the Development of the Cognitive Skills associated with the achievement of the proposed Objectives." (DEMRE, 2012d, p.5) The design of the specifications for this test assumed ordering the analyzed aspects in tables by level, warning that "only those Fundamental Objectives proposed in function of pedagogically measurable learning were considered, in accordance with the Cognitive Skills" referred to in Bloom's Taxonomy, which is used for the design of test specifications (p.31).

The theoretical framework for the Mathematics test (DEMRE, 2010a, p.4) indicates that "each item forming part of the PSU®−MAT is associated to an OF, since these are defined as

the competencies or capacities that the students must achieve when finishing the different levels of high school education." In this test, the skills measured have been defined in the OFs and are described as "standardizable procedures," corresponding to the methods and procedures for carrying out estimations, calculations, and the application of algorithms; "problem solving," related to data analysis and the proposal of answer alternatives; and structuring and generalization of the mathematical concepts, in relation to the identification of patterns and regularities and to the establishment of logical relations between concepts. In the review of the curriculum, contents were found that "aim towards the knowledge of the history of Mathematics and to the use of the calculator or computer software applied to the different contents proposed by the Curricular Framework," which "due to the nature of the items of this test" were identified as Not Measurable (NM) (page 6).

Page 14 of the theoretical framework for the language test indicates that the current test structure is formed by three sections: (1) knowledge of basic concepts and general language and communication skills, (2) text production indicators, and (3) reading comprehension and context vocabulary. It indicates that "the questions of each section are based upon the CMOs proceeding from the three subject axes that form the curricular framework of the subsector: Spanish Language, Literature, and Mass Communications Media." Later in the same document, the methodology for the production of the test specifications matrix describes the purpose for reviewing the curricular proposal of the language area (or subsector) in the high school education level "in accordance with a set of categories that endorse its transference towards a measuring instrument of the CMOs and which integrates the competencies, skills, and abilities represented by the OFs present in the current Curricular Framework for the subsector" (DEMRE, 2012c, p.20).

Finally, the History and Social Sciences test "is structured over a double entry curricular reference matrix" (contents and cognitive skills) (DEMRE, 2010c, p.29-30), among whose characteristics we may count having three subject axes, and that include cognitive skills associated with the pedagogical actions established through the OFs of the curricular framework. It also indicates that the item distribution between the three subject axes is coherent with the CMOs of the national curriculum.

In other frames of reference, it is stated that there are subjects that are not susceptible to assessment by means of paper-and-pencil tests and therefore are not incorporated into the test specifications. Globally, based upon the interview sessions with the DEMRE's technical team, the relevance of the curricular subjects has been considered in light of the formal requirements for admission into higher education. Based upon all of those factors, it may be estimated that approximately 80% of the curriculum is measured with the tests (MINEDUC, 2011).

In the tables of specifications for the tests, which form part of the theoretical frameworks, these contents are identified by the acronym NE (not evaluable). However, this information is only for internal circulation within each committee, and what the public generally acknowledges as the PSU referent is the national curriculum in its comprehensive form.

Nevertheless, during the interviews it was also pointed out that, since the administered tests are published, high school teachers, or any person interested in the curricular coverage of the PSU, may analyze the contents that are being evaluated and even further may determine the proportion of items of the test that fall into each subject axis. This analysis could then inform training practices vis-à-vis the relevance of teaching various aspects of curriculum content in the classroom. In any case, the team interviewed maintains that the committees are interested in evaluating what is relevant for higher education and

what in reality is part of the training processes of high school education. To that end, there are instances when items are excluded from a test when the reviewers determine that they are not relevant to either higher education or high school education. However, this type of decision is not the equivalent of avoiding the evaluation of certain content altogether; rather, it is an acknowledgement that other forms of measurement (viz., developing better items) is necessary. This entire process ensures that the table of specifications remains unaltered throughout the applications, as far as the percentage distribution of questions in the different axes considered.

As part of its duty, each committee writes and adjusts the theoretical framework of its respective test in accordance with the certain necessary technical considerations. Currently, among the important considerations that DEMRE must address is the 2009 curricular reform. This reform assumes changes to the test frameworks starting in 2013–2014, when the high school graduates, that is, the applicants to higher education, shall all be students trained under this new 2009 curriculum. Up to this date, the curricular frameworks of the PSU tests have been inspired by the curriculum in force between 1998 and 2008, attending to the fact that the applicants that have been taking the tests have been trained under that curriculum. Given the national curriculum changes in 2009, the PSU is undergoing a transition process which has driven DEMRE to revise henceforth all theoretical frameworks in order to ensure that the PSU continues to be pertinent and relevant toward the process of selecting university applicants that have a different training process from those that were trained under the 1998–2008 curriculum.

Regarding the specific theoretical frameworks of each test, these documents have similar structures to the extent that they include detailed lists of CMOs and OFs of the national curriculum for the specific area in each one of the four years of high school education. They describe the contents and cognitive skills that shall be the object of assessment in the respective tests, separating this information by educational year as well. They include the test specifications table itself, which describes the number of questions to construct for each cell of the table and, in the cases where there are different types of items, the technical indications regarding the construction of each type of question. Furthermore, the frameworks present at least one example of item formatting, which shall help for the subsequent revision, qualification and storage of items in the bank.

Committee members in charge of the tests of the PSU battery are professors of their respective disciplines, commonly with one or more postgraduate degrees.

The experience of each committee is measured by means of two main criteria: on the one hand, by their permanence in the committee (currently, half of the members of the committees have five or more years in DEMRE), and, on the other, by the diversity in postgraduate specializations that they have obtained—education, curriculum and assessment and didactics, among others. Also, most of the members of each committee have teaching experience in high school education and / or higher education in their specialization areas and several of them are also professional advisors in such areas (DEMRE, 2011).

Apart from the elaboration and writing of the theoretical frameworks and test specifications, the UCP committees are also responsible for training item writers and for revising the work produced by the commissions that have been placed in charge of item construction. During this process, each committee is responsible for making the theoretical framework characteristics of the test known and for training the members of the commission with respect to the general techniques and specific guidelines for item elaboration.

The technical documentation of the PSU conceptual frameworks is managed by the respective committees of the DEMRE UCP; there is no formal documentation describing protocols for the management and updating of these documents. The content coverage of PSU assessments and related information is available from official PSU web site and newspapers.

EVALUATION

The technical teams in charge of writing the specifications have academic training and professional experience that support their engagement in the development of test specifications. The test frameworks and the specifications are the product of a rigorous and thorough curricular analysis which has also contemplated the demands of higher education in order to pose a serious and objective assessment proposal. However, neither in the framework nor the specifications construction is there evidence of outside DEMRE participation of experts in the review and analysis of the pertinence and relevance of these specifications for the purposes of the test and for the interest of the population under assessment. Even if during the processes of item construction there has been participation of university academicians who are otherwise not part of the process, such participation does not imply a critical stance vis-à-vis the frameworks or specifications as documents that guide the construction of the instruments. Given that this is a nationally administered test, it would be advisable to include more outside views that would bring to bear different perspectives (either different levels of training levels or differing disciplinary specialties) that would judge the adequacy of the test specification as well as the test items that are thereby developed.

Furthermore, though the test specifications include guidelines for item development, they could include more detailed examples of the different sectors of the table. They could also incorporate more analysis and observations with respect to possible limitations regarding any unexpected or undesired results among the target populations. Even though the specifications include examples of items and descriptions of the item formats that are used, these are insufficient to illustrate the different item formats and different categories of the assessment objects.

Additionally, it is evident that the technical team is clear on the fact that these documents will have to be updated in the near future (since the curriculum change will become effective as of 2013) but currently there is no evidence of the existence of a clear plan as to how this will take place nor of the studies that are being carried out or that will take place in the immediate future to support these modifications.

Even though the theoretical framework documents and test specifications for each test exist, the depth and detail of the amount of information is not well standardized across the tests and their committees. There would be a higher level of systematization and technical rigor if all frameworks and specifications were to be elaborated under similar parameters.

Table 2 shows a summary evaluation for PSU quality, security and confidentiality standards regarding the development of items and tests. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with the purpose of identifying finer grains of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 2: Summary Evaluation of PSU Quality, Security and Confidentiality Standards regarding the Development of Items and Tests

Facet 1: Item Development Framework and Specifications	
1. Describe the PSU process for development of frameworks and specifications / guidelines for item development and writing which appear in the test.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. How are the frameworks and specifications updated and when? • Which is the updating process? • What types of considerations are given to changes in the target population of the test? What types of considerations are given to changes in the purpose of the test (for example, NRT and CRT)? • How does research inform the updating process? • How are the intentional and unintentional consequences assessed? What is the time framework allowed for the modifications to take place?	C (3.1)
b. Who writes (or determines) the specifications and what are their aptitudes?	F
c. Which is the process to examine and review frameworks and specifications? • How different would the process be if research were to take	E

place as part of the framework examinations and revisions? • Is there a pre-established assessment process of the DEMRE process followed regarding the development of frameworks and specifications / guidelines for the development and writing of test items? • Which are the most important characteristics of the assessment process? What type of assessment criteria is utilized? • Who carries out the assessment and writes the reports? • What type of use is given to the assessment results?	
d. Is there sufficient detail? • Why is there more trust placed in the opinion of a content expert?	E
e. Do the specifications / guidelines for item development detail the particular needs of the PSU testing program? (That is, are they specific?) • Do they provide a context? • Do they register relevant similarities and differences with respect to the previous years? • Do they analyze unexpected potential results due to changes in the target populations, meanings and score uses?	E
f. Are there example or model items included or available in the specifications?	E
g. Are the specifications easily accessible for the item development team requiring access? • Who carries out the documentation maintenance?	C (3.2)

RECOMMENDATIONS

1. In order to provide for a more rigorous test design, we recommend that the documents state clearly the purposes and uses of test scores and the nature of decisions to be made with the scores (e.g., norm reference/criterion reference), the intended population of test takers, the definition of domain of interest, and the processes outlining the development of test frameworks and test specifications. Because the PSU is often referred as a battery, it is pertinent to expect a consolidated test specifications document for the tests it comprises. The international evaluation committee would like to stress the importance of developing such documentation with an intended user in mind.
2. In relation to the rigor that the definition and explanation of the dominion of the test must have, the recommendation is to advance studies on the effect that the decision to approach the test with a priority on the CMOs of grades 1 and 2 in high school education may have, as well as determining the effects of the fact that the test may have greater weight for the Scientific–Humanistic curricular branch than for the Technical-Professional one. It would be desirable to foster the materialization of the policies referred to during the interviews with the technical team in the sense of including in the PSU aspects of the Technical–Professional curricular branch, or to study alternatives for the equitable assessment of the populations formed under both curricular branches.
3. According to the need to include an expert external review of the specifications, it is recommended that the theoretical frameworks and the specifications of the tests be submitted for validation by readers from outside the committee members who do not have the function of constructing items. This independence would allow for feedback

concerning aspects such as the adequacy of the content coverage (axes), as well as the decision concerning the inclusion or noninclusion of CMOs in the same. Part of the framework validation process should be carried out with the participation of experts and with pertinent documentation. The description of the expertise of reviewers is important to collect and report within PSU test framework documentations. This documentation should cover descriptions and examples of materials evaluated, with particular emphasis on test domain, test use, and intended populations of test takers. Finally, we strongly recommend developing, administering, and summarizing participants' answers to a survey after the evaluation exercise and using their recommendations for improving the evaluation process for subsequent years.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2010a). *Descripción técnica de la prueba de matemática.* Santiago: Universidad de Chile.

DEMRE. (2010b). *Entrevistas con el equipo técnico. Transcripción No. 2. Chile – 2012*. Santiago: Universidad de Chile.

DEMRE. (2010c). *Marco teórico prueba de selección universitaria historia y ciencias sociales.* Santiago: Universidad de Chile.

DEMRE. (2010d). *Prueba de selección universitaria (PSU): Antecedentes y especificaciones técnicas.* Santiago: Universidad de Chile.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2012a). *Actualización marco teórico ciencias naturales (física).* Santiago: Universidad de Chile.

DEMRE. (2012b). *Marco teórico de la prueba de selección universitaria (PSU) del sector ciencias naturales subsector de biología*.

DEMRE. (2012c). *Marco teórico prueba de lenguaje y comunicación.* Santiago: Universidad de Chile.

DEMRE. (2012d). *Marco teórico PSU-ciencias-química.* Santiago: Universidad de Chile.

MINEDUC. (2011). *Aprueba bases administrativas, bases técnicas y anexos de licitación pública, sobre servicio de evaluación de la Prueba de Selección Universitaria (PSU)* (ID Nº 592-44-LP11). Santiago, Chile: Autor.

Objective 1.1.a. Facet 2. Process for item writer selection and assessment

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for selecting item writers. A framework for evaluating PSU approaches for selecting item writers is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.6

> The type of items, the response formats, scoring procedures, and test administration procedure should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that that intended inferences from test scores are equally valid for member of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judge should also be documented. (p. 44)

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items, from the item pool should be documented. If the items were classified into different categories or subsets according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented. (p. 44)

GENERAL DESCRIPTION OF THE PROCESS

Concerning item writing, DEMRE organizes item construction commissions for each area assessed, which operate under the coordination of the respective DEMRE committee. Even though each construction commission develops its own particular dynamics, based on factors such as the volume of the items to construct or the degree of experience in item construction of the commission members, the general process for the selection and training of writers is similar.

As background to the writer selection process, each test committee, composed of five DEMRE professionals, studies the needs or requirements for item construction for the subsequent PSU applications. They establish the need to form a construction commission and a summons is issued via the DEMRE web page, which consists of a call for applications for the position of "PSU Item Writer" ("Constructor de Ítemes PSU"). Because the process is open, each committee receives numerous résumés from applicants aspiring to be part of this commission. The respective committees must filter the applications in accordance with two basic requirements: (1) graduate studies in the field and (2) teaching experience in the field. The teaching experience required is at least three to five years experience. There are additional criteria, such as having taken courses in technical training regarding item writing, or in psychometric aspects

in general, and being proficient in the use of computer tools, even though these are not necessarily definitive for the selection process due to the fact that the commission shall assume the responsibility of training the members of the committee in the required technical aspects of item construction. The committees may apply additional selection criteria such as geographic diversity or the professional specialization of the constructors; however, these particular criteria cannot always be applied because of certain practical difficulties, such as the displacement of the constructors or the availability of applicants with the required specializations.

Once filtering of résumés of the applicants has taken place, prospective members of a commission, especially new applicants, are called for a personal interview, after which the respective committees decide on the commissions.

In the particular case of the Language and Communication test, the selection involves the applicants taking a test on curricular proficiency. This test evaluates their knowledge of the curricular framework, their knowledge of the axes or thematic fields of the curriculum, and the cognitive skills and competencies entailed in item formulation. Those teachers approved at this stage go on to form part of the yearlong course taught on item elaboration, through the Continuing Education Program for Teachers (PEC), with support from the Faculty of Philosophy and Humanities of the Universidad de Chile. Once this stage is completed, the selected teachers become part of the PSU Language Test Constructor team.

To receive a Mathematics and Chemistry commission, preselected teachers must go through a workshop on item construction and analysis taught by the same committee. The final selection is based on the performance of the applicants in this workshop. DEMRE requires educational institutions to which the authors are linked to endorse the authors prior to their participation on the commissions, thus ensuring that the constructors will be available to participate throughout the entire construction process.

In general the selection process culminates with the forming of one or more constructor commissions, each one consisting of five members: two academicians representing the higher educational level and two high school teachers, who together are led by a member of the respective DEMRE committee. This composite structure of the commissions allows specific roles to be carried out by the commission members, by which a synergy of competencies is expected, enabling a higher item quality. The two academicians, who are higher education teachers, preferably of the first academic semester, shall act within the committee as president of the commission and as technical advisor of same, respectively. The knowledge that these academicians have of the university requirements with respect to disciplinary knowledge in their field, as well as, given their academic exercise, their current level in the theoretical progress of the discipline, constitute their contribution regarding the analysis of item pertinence and relevance. For their part, the two high school education teachers mainly contribute their knowledge of the high school education National Curriculum and their close knowledge of the population taking the PSU, which offers security regarding the pertinence of the questions for the population. The DEMRE representative, who is one of the regular committee members of the respective test, is responsible for coordinating the group in the discussions during the item revision workshops, as well as for the administrative functions corresponding to the commission meetings or, in case it is required, of obtaining an additional expert to sort out controversies within the commission around the pertinence of a certain item.

Each DEMRE committee decides whether to form a single commission for item construction or to form subcommissions, which may mean that more than one constructor commission will be working simultaneously under the coordination of the same committee. The decision to subdivide the work is based on the volume of questions that must be constructed and, in general, the criteria for the division into subcommittees are related to the subject subdivisions of each test.

The commissions undergo a training process that has three components: administrative information on the functioning of DEMRE and the dynamics of the commissions, technical information on item elaboration and information on the handling of specialized software for question construction.

The training of the commissions varies as a function of the degree of experience of their members. The new constructors receive a more specific orientation that provides them information about the DEMRE structure, how this structure operates, and the work goals during the year. In contrast, the experienced constructors, who already know the dynamics of DEMRE and its committees, are presented with the existing construction requirements and, if necessary, with the updated procedures. In general, an important part of the training for all, new as well as experienced constructors, consists of the analysis of items that were piloted but which did not pass the technical criteria to form part of the bank, with the purpose of recognizing the more frequent construction faults and to avoid them in the construction being started.

Even though the dynamics of the constructor induction and training process are different between the different DEMRE committees, the central aspects that constitute this process are the following:

- Organizational structure of DEMRE
- Purposes of the PSU
- Item elaboration process followed by DEMRE
- Types of items required to be constructed
- Formal technical criteria for item construction
- Item approval and rejection process

Besides the training provided through meetings or workshops, the item writers are provided documentation containing the theoretical frameworks of the test, the specifications tables and criteria for item construction. These documents, on which the constructors will rely during their work, are analyzed in detail as part of the orientation.

All of the commissions also undergo training in the process of handling the "Safe Question" software, which is used for item construction. Some are trained in specialized software, such as Freehand (in the case of Mathematics) and ChemDraw (in the case of Chemistry).

Once the commissions or subcommissions have received the information and training provided, the assignment of item construction tasks takes place. Each commission member is assigned responsibility for the construction of a certain number of items related to specific parts of the specifications table. At home, each constructor individually elaborates the questions and then attends DEMRE for item revision

meetings. It must be pointed out that one of the obligations of the item constructors is to maintain confidentiality concerning the items they develop, which is the reason why each one signs a document pledging to keep this information confidential.

During the development of the commission's activities and upon terminating them, constructor assessment processes are carried out. In general terms, this assessment construction takes place in accordance with the function entrusted to each member of the commission but with the following as its central aspects: the efficiency of the work carried out (number of items made with respect to the number of items approved), the creativity in item construction and the level of participation and contribution in revision discussions. Even though constructors are expected to meet the criterion of a 75% item approval rate, the DEMRE committees acknowledge that this goal is achieved as experience with construction is gained. This is the reason why the criteria of creativity and participation in the revision discussions may have more weight in the assessment of the performance of a constructor. Based upon these criteria, the committee members periodically provide feedback to the constructors, calling attention to the mistakes that are being detected in such a way as to give the constructor the opportunity to correct them and thus remain in the process.

The quantity of items constructed in the item bank by each member of a commission is registered in monthly reports, which are evaluated once a year by the respective DEMRE committee. Together, the lead DEMRE committee member and the academician who performed the role of president of the item construction commission assess the number of rejected items from each constructor, as well as their participation in the process carried out. Based upon this review, they decide to call back individual constructors to form part of the next construction commission. Regardless of whether the constructor shall be called back or not, the constructors are offered a verbal assessment of the work performed. In essence, the assessment of the item writers is predominantly a verbal process with the exception of the monthly written reports, which register the rate at which questions were successfully entered into the bank by each constructor.

EVALUATION

The technical team in charge of coordinating the selection and appraisal of the item writers has the adequate academic and technical information to carry out these labors. The selection process is characterized by being open, which contributes to the transparency of the process and to the variability of the writer teams. Also, it is acknowledged that there is an important effort in achieving diverse construction teams representing the interests of the test users, that is, care is taken that the teams represent high school education and higher education institutions. Though Standards 3.6 and 3.7 do not explicitly require documentation of recruiting procedures, formal documentation for choosing item writers and test constructors would provide evidence that supports the consequential validity of the PSU.

Generally, even though the recruitment of writers could continue using the current mechanism of open call through the DEMRE web site, it would be better if it were complemented with personalized invitations, open calls through municipal authorities, etc., in order to form writer teams that are diverse not only with respect to their areas of expertise, length of their teaching experience, and their methodological orientations, but also with their institutional representativeness (e.g., Scientific-Humanistic versus Technical-Professional curricular branches).

In accordance with standard 3.6, (AERA, APA, & NCME, 1999), the qualifications, the experience and the demographic characteristics of the item writers should be documented. The validity of the test inferences can be affected by the diversity of the item writer population. Even though the open call on the DEMRE web site for item writers does not restrict who can participate, such a mechanism does not guarantee the representativeness of a team of item writers.[1]

Table 3 shows a summary evaluation for PSU scales and standardized scores. The purpose of the table is to provide a high level snapshot of expert evaluation of the holistic evaluation of the second facet of the evaluative objective (1.1.a. Process for Item Writer Selection and Assessment).

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parenthesis.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment

---

[1] For example, geographic representativeness is a matter of importance since it concerns a nationally administered test, which affects the students of all regions of the country. The training processes in rural regions or regions that are far from the capital city tend to place particular emphasis on different aspects of the official curriculum and established standards. In the educational context, it is clear that the official curriculum is one thing and the curriculum actually implemented in the classroom may not be exactly the same one, despite efforts made by educators and administrators to pay close attention to the official curriculum. In this sense, to include into the construction process teachers from distant regions would help substantiate to a greater extent that the sample of test questions is germane to the diversity of contexts and conditions often found among schools from all regions of the country. It would also facilitate the dissemination of the evaluation process of the PSU. For as teachers come to know more about the test construction process, they will help disseminate technical and academic aspects of the PSU's construction to their colleagues and students. Moreover, it would also facilitate and reinforce the importance of aligning instruction to expectations stated on domains adopted for the test.

standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 3: Summary Evaluation of PSU Process for Item Writer Selection and Assessment

Facet 2: Item writer training and selection		
1.	Describe the processes followed for the selection and training of item writers.	
2.	Follow-up questions (in case they are necessary and if applicable)	Rating
	a. Are there pre-established criteria regarding the recruitment and selection of item writers? • Which are the most important characteristics of the criteria? • Once elected, how long do the item writers remain? • Are the item writers demographically diverse? • Are the item writers professionally diverse? • Are the item writers geographically diverse?	E
	b. Is there a pre-established process for the evaluation of the item writers? • What are the most important characteristics of the criteria for evaluation? • How is the performance of the item writers assessed?	E
	c. Is there a pre-established assessment process of the DEMRE process for the selection and training of item writers? • Which are the most important characteristics of the assessment process? • What types of assessment criteria are utilized? • Who carries out the assessment and writes the reports? • What type of use is given to the assessment results?	E
	d. Do(es) the item writer(s) have a deep knowledge of the current curricula and of the disciplinary content? • How is it assessed?	E
	e. Do(es) the item writer(s) have a complete knowledge about the current students (that is, experience in the Technical-Professional and Scientific-Humanistic)?	E
	f. Is (are) the item writers trained in the item writing principles? • Does the training acknowledge the difference between NRT vs. CRT?	C (3.6, 3.7)
	g. Is (are) the item writer(s) trained at some level in psychometric principles?	C (3.6)
	h. Do(es) the item writer(s) receive training in the use of the specifications?	E

RECOMMENDATIONS

1. We recommend including much more explicit documentation regarding the participation of item writers, capturing educational specialty levels, length of teaching experience and region of origin in the country. The use of information technologies could be used for this purpose, enabling the participation at a distance of item writers for whom reaching the capital city of the country is difficult.
2. Item writer training and certification processes should occur to ensure the quality of the tests and the validity of the evaluation process. The participants should respond to challenging criteria that formalize the process so that all of the item writers engage in the work under similar conditions. They should have a basic grasp of the discipline in which they are to construct questions; a basic knowledge of the purposes, theoretical framework, and the test specifications; and, finally, the basic technical knowledge concerning the process of test item construction. Each committee allocates a different amount of time to the training process and the work times, and the emphasis on the subjects treated in the induction meetings for item writers are not standardized. In order to ensure an adequate competency level for an item writer, it is necessary to standardize the training process with respect to the objectives, topics, times, materials and the rest of the resources, as well as process control and evaluation mechanisms. It is necessary to give each writer the opportunity to develop his or her item writing skills and to receive timely feedback on the items before that writer begins writing items for the pilot. The Language and Communication committee follows the most formalized item development process. It would be worthwhile to generalize this process to the remaining content areas so as to ensure uniform training of the commission members.
3. We also recommend implementing a certification system for item writers completing the formal training process in order to develop a pool of certified item writers. This could be done by DEMRE directly, or an institution that may offer training to writers. In any case, the training process should be designed to include:

    - Verification that the item writers have a grasp of the discipline in which the items shall be developed, emphasizing the assessment content
    - Theoretical training in technical aspects of the test design, such as specification matrices, item construction rules, scales and results reporting, and measurement concepts such as validity and reliability
    - Training in item writing through workshops with detailed feedback concerning the individual mistakes in construction
    - Training in item analysis, including conceptual aspects of psychometric indicators and practical interpretation exercises of same, emphasizing the relationship between item construction characteristics and the indicators obtained
    - Training with respect to the particular purposes of the PSU, its background and uses

    The training shall include participant evaluations to verify a minimum level of understanding of the topics as well as the quality of the items produced.

    The final purpose shall be to grant each participant status as a certified item writer. Also, depending upon the changes introduced in the test or in its development procedures, periodic update processes should be contemplated for

the certified item writers. Considering the quantity and characteristics of the topics recommended for item writer certification, we would estimate that a complete course of item writer training such as the one described above would take at least 30 to 40 hours.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

Objective 1.1.a. Facet 3. Commissioning — Item writing process

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for writing items. A framework for evaluating PSU approaches for writing items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.6

> The type of items, the response formats, scoring procedures, and test administration procedure should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that that intended inferences from test scores are equally valid for member of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judge should also be documented. (p. 44)

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items, from the item pool should be documented. If the items were classified into different categories or subsets according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented. (p. 44)

GENERAL PROCESS DESCRIPTION

Item development for pilot and definitive tests is the responsibility of the Test Construction Unit (UCP), represented in the committees responsible for each one of the PSU tests. These committees organize item writer selection processes for the formation of item writer commissions that are responsible for writing the required items.

Item writing is carried out by each member of the constructing commission individually in accordance with the assignment given to each member of the commission. In the Language test, this assignment takes place by means of a construction road map, according to the annual review requirements for questions from the item bank. In the remaining areas, an assignment takes place week by week, in what is known as "construction map," also in accordance with the needs detected in the bank.

The UCP leads are responsible for selecting valid and reliable item writers. The process to select the item writers is described as part of Facet 2 of this objective. The item writers are recruited to write between three and six items each week and to attend weekly meeting at DEMRE to review the items. This process normally takes place between March and August each year. The item writers are solely responsible for respecting copyrights of source documents consulted during the item development process.

Each item must be constructed using a form in which, aside from the item itself, the author must register the item's descriptive aspects (theme, key, item difficulty and curricular aspects under assessment), which are used as criteria in order to carry out the analysis of each item.

Concerning item construction, the authors receive software called "Safe Question," which is personalized for each author with a password login. This software includes a word processor, which makes it possible to insert graphic elements in a way such that the authors do not have to resort to using other software to finish writing their items. In the case of texts used for the construction of several questions (such as in the case of Language tests), the software allows the inclusion of these as independent questions and the item bank system establishes the corresponding link between each text and its derived items.

The Safe Question software seeks secure the confidentiality of the items constructed since a file encryption mechanism operates automatically, which prevents them from being opened by any other different software. The authors work on the items assigned to them in their own homes and bring the files of their files to the revision meetings where they are decrypted by the DEMRE computers. The software automatically eliminates the original version of the item from the storage device that the author has taken his/her items to the meeting.

With respect to adaptations of the PSU questions for populations with some slight disability, it is specified that a larger-sized version of the test is to be printed, since it concerns persons with slight medically certified visual disabilities. Disabled applicants are provided special treatment when taking the PSU, the sole purpose of which is to enable their participation in the Admissions Process under fair conditions without any implied commitment for their acceptance on the part of the Universities, which reserve to themselves the right to grant admission as they deem appropriate. People suffering mental illnesses are exempted from this special attention, as well as people evidencing deafness or stammering, since they are not impeded from taking the tests, inasmuch as the instructions written; therefore, their compatibility or incompatibility with the incorporation into a career or program is exclusively subject to the regulations of each university.

In order to exercise the right to this attention, those interested shall register themselves through the normal process, and submit a written request, before October 30 of the current year, addressed to the *Dirección del Departamento de Evaluación, Medición y Registro Educacional (Avda. José Pedro Alessandri 685, Ñuñoa, Santiago*). The request shall include:

1. Identification of the applicant
2. Clear exposition of the reason for the request
3. The following documents shall be attached:

    - A medical certificate, issued by the treating physician, determining the pathology affecting the applicant and clearly indicating the degree of disability presented. At the moment of its submittal to DEMRE, the certificate shall not be more than three months old starting from its issuing date.
    - A photocopy of the certificate issued by the respective *Comisión de Medicina Preventiva e Invalidez (COMPIN)*, or a copy of the personal ID

card or certificate issued by the *Registro Nacional de Discapacidad*, should it correspond.
- In the case of visual pathologies, an examination of visual acuteness shall be enclosed. In case it is feasible, adding an examination of the back of the eye (retina, etc.) is also suggested.

Should the application not be in compliance with the requested requirements, it shall be flatly rejected and for all purposes shall be considered as not submitted.

The final resolution regarding the conditions in which the tests shall be implemented is the competency of the Medical Service of the Universidad de Chile. To tend to cases in these regions, the collaboration of the medical services belonging to other universities belonging to the *Consejo de Rectores* may be requested. In the event that the attached background information happens to be insufficient and an expert examination of the applicant is requested on the part of the Medical Service practicing the assessment, these costs shall be defrayed by the applicant. The decision issued by this institution is final and may not be appealed.

Concerning the treatment of blind people, it is specified that these are the only people who are disqualified from taking the tests. This is due to the fact that even if they were aided in reading, it would be impossible to represent to them the visual elements appearing in several tests. For this reason, their entry into Higher University Studies must take place through the special admissions processes to those universities that contemplate this alternative and into career studies which are compatible with their conditions. Persons who find themselves in this situation and who have their High School licenses or are currently attending their last high school year may request the collaboration of DEMRE directly, or through the admissions offices, who will guide them in their university applications through special admissions. For that purpose, they shall forward a note identifying themselves and enclose the required documents (consulted in DEMRE (2010, 9 de Diciembre).

Currently, SENADIS, the National Disability Service, is carrying out a joint project with Universidad Católica and Universidad de Chile for the development of special tests for blind people as well as for deaf mute people, but it still is in its initial phases and currently there is no special test for these populations.

EVALUATION

The item writing process seems to be given an adequate time period, which is positive for the process since it ensures that an adequate time for analysis and adjustment may be dedicated to each item. The item developers write the items from the respective institutions where they work, which allows for their availability when developing items and attending item review meetings. The number of items assigned to each item writer is reasonable. However, the review process could be improved if the coordinator of each commission performs an item review before they are taken to the joint review meetings, making use of standardized checklists for compliance with basic item quality criteria. This would facilitate the feedback to the writers identifying which item aspects do not comply with the established standard and would optimize the time of the review meetings.

The template used for item writing has key information for characterizing the constructed item, even though it could include more precise information on, for example, the person in the commission in charge of performing the first item review (if

so designated), the application restrictions it would have over particular populations (disabled people, for example), the possible incompatibilities with other elements constructed during the process, etc. That is, it would be information which may later be fed into the item bank and which would facilitate subsequent selection processes for test assembly, as well as follow up tasks to institutional processes (quality audits). It is advisable to implement policies on adaptation of the tests to be applied to populations with different disabilities to better capture their actual level of test performance and add elements of fairness to the process.

Table 4 shows a summary evaluation for PSU item writing process / commissioning. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with the purpose of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parenthesis.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 4: Summary Evaluation of PSU Item Writing Process / Commissioning

Facet 3: Commissioning — Item Writing Process	
1. Describe the processes followed for the PSU item writing process / commissioning.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Does the process allow for feedback during the initial construction of the draft items? • Who produces changes in the initial draft versions of the items? • Are there documented criteria for conducting changes in the initial draft versions of the items? • Why do you think that a criterion is or not necessary? • When was the criterion developed and when did the last update take place? o I have not heard of changes that may be introduced into draft versions for acknowledging the diversification of curricular branches (Scientific-Humanistic vs. Technical-Professional). Would you mind clarifying this?	C (3.6; 3.7)
b. What is the availability for item writers to ask questions?	E
c. How are the items commissioned? • Who determines the quantity of items needed? • What process is followed? • Are the reading passages commissioned separately from the items? • How many items are written per passage?	E
d. How much time is allotted for item writing?	E
e. What type of item template information (metadata) is supplied? • How is it that the item writer(s) obtain(s) this information?	E
f. What kind of electronic system is used for drafting items?	E
g. What considerations are given to accommodations (Universal Design)? • How much research has informed the decisions on accommodations?	C (3.6; 3.7)

RECOMMENDATIONS

1. We recommend performing systematic checks on the assumptions that the item writers are following the principles of confidentiality and copyrights. These checks would entail having senior content staff from DEMRE perform random checks on test content item and art against major sources of copyrighted materials.
2. As stated in Standard 3.7, "[t]he procedures used to develop, review, and try out items, and to select items, from the item pool should be documented." We recommend clear and transparent documentation of the process for surveying the item bank to identify the quantity of items to be commissioned. Along this line, we would like to recommend specifying more precisely how item writers are to be selected.
3. We recommend studies that identify the characteristics of the items that can be adapted during item development (such as graphic materials included, letter fonts, diagramming and editing aspects in general, among others) which may make reading of same easier for the population with special disabilities. Although the procedure established by the UCP to increase the font size and graphics for the visually impaired is relevant and represents an important element in ensuring equity in the assessment process, there is still a need to explore alternative mechanisms, both in the actual construction of testing and in the implementation process to ensure more equitable conditions for all students. For example, what are the local physical conditions of administration or the distance travelled by disabled people to the administration sites? It is also important to investigate whether the granted accommodations match those used in the classroom. Intended accommodations, when they are not aligned to classroom conditions, have the potential of introducing construct irrelevant variance instead of removing it.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2010, 9 de Diciembre). *Recomendaciones para rendir el examen.* Proceso de Admisión 2011, El Mercurio. Documento No. 26 Serie DEMRE – Universidad de Chile.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.a. Facet 4. Process for item revision and approval

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for revising and approving items. A framework for evaluating PSU approaches for revising and approving items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and / or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented. (p. 44)

GENERAL PROCESS DESCRIPTION

Item approval and revision is the responsibility of the item construction commissions which depend on the committees in charge of tests within the UCP of DEMRE. These revisions take place during working sessions carried out in the DEMRE facilities. The revision process happens simultaneously with that of construction; that is, the items being written by the authors are being reviewed, in a process which may last some 16 to 20 weeks, usually between March and August every year.

The revision commissions, which are the same as the construction commissions, usually consist of five members. Two members are high school teachers in the field to be assessed, another two are higher education academicians, also trained in the assessment field, and all of them are coordinated by a member of the respective test committee, a DEMRE professional. The two academicians fulfill the role of president and expert advisor of the commission, respectively, and, according to their profile, they are expected to contribute from their specialized knowledge of their field as well as from their familiarity with the requirements that the university poses in that field on the students entering the first semester. For their part, the high school teachers contribute from their knowledge of the national curriculum of high school education and their practical knowledge regarding that training level present in the classroom. The DEMRE professional, a committee member, aside from coordinating the commission's work, must contribute the technical know-how on the respective test and on the technique for item elaboration. Due to the fact that it is the same commission members who construct and review the items, the selection of the reviewers reflects the same basic criteria applied in the recruitment of writers and in this sense ensures that the reviewers have degrees and postgraduate degrees in the field, as well as teaching experience at the respective level. On the other hand, even though it is intended, the geographic diversity of the participants in the commission is not guaranteed due to difficulties in maintaining continuing participation of persons from regions other than the capital city.

During each commission's review sessions, every constructor reviews between three and five items which have been allotted previously. Each session for item revision is projected to last from three to three and a half hours, during which around 15 to 20 items are reviewed. The questions proposed by each constructor are read collectively in the commission and are adjusted in conformity with item pertinence with respect to the assessment objectives of the test, for the national curriculum, on the mastery over the item subject in high school classrooms, and on the relevance of the item regarding entry into the university among other basic analysis criteria.

In general, for all commissions there are two broad categories for the review of items: the qualitative aspects and the quantitative aspects. The qualitative aspects refer to the construction itself, which uses the same test construction template utilized during item writing: subject and assessed cognitive skill, correct answer, item type, etc. The quantitative aspects refer to the analysis and interpretation of the item statistical analysis. The quantitative analysis will be carried out once the items have gone through a pilot application.

After the item revision and adjustment on the part of the commission members, the items go through a second review, this time by the DEMRE committee, who has the task of approving (or rejecting) the items. This work includes the participation of the person who has carried out the function of commission president as well as of the commission's technical advisor. In this phase, the member of the DEMRE committee who coordinated the group of constructors presents the questions approved in the commission, listens to the observations of each one of the members of the committee, and decides (with the full committee) upon the necessary adjustments to be made on the items to be left in the final version. If necessary, the commission president or the committee member that has managed the group may decide on the advisability of calling upon an additional expert, who has not participated in the revision process, to elucidate aspects of the items which may have been left unresolved in the commission discussions.

The result of the final item quality assessment in the committee is expressed, regarding the Language area, through a scale that has five categories: rejected, insufficient, sufficient, good and optimum. In other areas, only two categories are used: approved or rejected.

Generally, all of the constructed items are entered into the item bank, but they take on a different "condition" according to the revision dynamics and the decision concerning approval or rejection taken on them. The final item versions with the approved condition are those that are used for assembling the test.

The test assembled in a first version is subjected to the revision of the commission president and of the commission's technical advisor, who may suggest item adjustments or replacement. The assembled test, before its final printing, receives one last revision by the committee members, especially with respect to grammatical and editing issues.

In the test construction and assembly process, as well as in the revision of previously assembled tests, the following criteria are taken as basic points of reference for the revision of the test:

- Type of question (according to the formats used by the UCP in the different areas)
- Curricular suitability
- Subject assessed
- Correspondence with the educational level for which it was elaborated
- Assessed cognitive process
- Estimation of difficulty
- Relevance and pertinence
- Correct Answer
- Quality and homogeneity of the response options
- Grammatical aspects

All of the suggested recommendations for adjustment made by the commission or by the outside experts are recorded in minutes in order for the respective committee members to analyze and consider prior to making final adjustments to the tests.

Once the items are approved for pilot test assembly, they will be analyzed quantitatively, which will inform decisions whether they will be adjusted in new ways, entirely rejected, or approved with no modifications. All of the constructed items, even those rejected, remain in the item bank with the label corresponding to their condition. (Please note, there are 15 possible condition labels, two of which are approved or rejected). These condition labels are useful because they can be used to show future item writers and reviewers how certain construction faults lead to item rejection.

EVALUATION

The existence of a systematic procedure for item review and approval is acknowledged. Even though the reviewer observations are documented in minutes, there is no evidence with respect to the existence of documentation containing the item quality standards expected for each of the tests. In fact, the qualification scale for item approval or rejection used in the different tests is a different scale, which is evidence of the different criteria or at least of the different degree of precision used in valuing item quality.

Since the item reviewers are themselves members of the commission and also members of the respective committee for each test, item review could be very much enriched by the participation of outside reviewers who have not participated in the construction process. These outside reviewers could issue unbiased judgments which may originate from having participated in the review meetings. A reviewer who knows his target population well and who knows his way of reading and interpreting that which is read could be of great use as an outside reader of the individual questions or, even better, of the assembled tests. He or she could offer an objective evaluation on aspects such as text clarity, their length, their pertinence and even on how interesting the questions included are.

If it was possible to include more than one of these reviewers, it would be ideal for them to be representatives of different types of institutions (public, private, rural, urban, emphasizing both curricular branches, etc.)

Table 5 shows a summary evaluation for the PSU item revision and approval process. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 5: Summary of the Item Revision and Approval Process

Facet 4: Item revision and approval process	
1. Describe the processes followed with respect to item revision and approval.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Which is the revision process with respect to item approval?	E
b. Are there specific types of revision procedures for item approval by means of outside committees? • Is there any reason in particular to avoid outside item revisions?	C (3.9)
c. How many items are reviewed typically?	E
d. What are the criteria for the selection of item reviewers which operate in your committee?	E
e. What is the composition of item reviewers or committees (for example, fairness and sensitivity)? • What is the geographical composition of your item revision committees? • What is the geographic representation of your item revision committees? • Which is the balance between teachers and faculty members in your item revision committees? • Which is the balance between the Scientific–Humanistic and Technical-Professional branches of your item revision committee? • Which is the balance between private, subsidized and municipality in your reviewing committee?	C (3.9)
f. What type of training is provided to the item revision committees?	E
g. What revision criteria are followed regarding item approval during the meetings of the item revision committee? • Could you please list the elements of said criteria?	C (3.9)
h. How are the recommendations of the item revision committee processes handled? • How are the items rejected by the revision process handled? • What happens with them?	E

RECOMMENDATIONS

1. We recommend increasing the efficiency of the initial draft item review process by involving the senior content staff from DEMRE prior to the full committee review. The purpose of this senior review is to check compliance to item specifications, e.g., the content relevance of items, the appropriateness of the items for different populations, and the application of editorial guidelines.
2. We recommend that DEMRE formalize the process of providing feedback to item writers in a clear and objective way. Such documentation would provide information for analyzing common errors during item writing and therefore guide future training of item writers and reviewers.
3. We recommend that the procedures used to develop, review, and try out items, and to select items, from the item pool should be documented.
4. We recommend using a panel of item reviewers that is independent of the panel of item developers. This item reviewer panel should consist of a group of

qualified item writers who did not participate in the development of the particular set of items under review.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.a. Facet 5. Authorship tool and item bank

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for authoring items. A framework for evaluating PSU approaches for authoring items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

STANDARD NOT IDENTIFIED.

GENERAL PROCESS DESCRIPTION

Once the item construction and revision process is completed, item writers and reviewers load the items constructed into the item bank platform.

To load the information corresponding to the revision session, different software tabs are used to enable the registration of the names of those participating in the session, the date when the session took place, the descriptive characteristics of the items, and, finally, the item condition. Generally, the condition that appears by default is that of "proposed," but the person loading the system may change that condition to "approved" or "rejected" or rather "experimental" or "official," or several other possible conditions within the system, depending on the decisions which may have been taken with respect to the item. The software allocates a unique number to each item in a way such that it remains unequivocally identified within the item bank.
Given the fact that all the parameters and characteristics of the items and the items themselves are found in the SAFE QUESTION module or program, all of this information can be loaded directly into the item bank.

In order to guarantee the security of item banking, the loading of the items along with their characteristics is done at specified times and only by the person from the UCP responsible for the test. The permission to load items is given strictly by the item bank administrator, who is the only person who has, as his/her role indicates, an administrator profile, but who cannot himself/herself access the items in the bank (this the only restriction to the item bank administrator). Access to the facilities where the bank is physically located is restricted to certain UCP personnel, some of whom have the physical key (hardware key) that provides access to certain restricted-access files. The item bank operates as an internal network in which the contents as well as the internal validation procedures are automatically encrypted in the system. Finally, all of the item bank information is placed on a safe server used exclusively for this purpose.

In general terms, the UCP is the unit responsible for the operational management of the item bank. Its professionals are responsible for administrating the access profiles to the items and to the tests. In cases where ongoing research is performed, certain professionals from the Studies and Research Unit also are provided access for the item bank in order to obtain the information needed for their research studies.

For populations with some degree of disability, such as a slight, medically certified visual disability, an enlarged version of the test must be printed. Disabled applicants are provided special treatment regarding taking the PSU. These accommodations are specifically targeted to allow these students to participate in the admissions process under fair conditions, without any implied commitment for their acceptance whatsoever

on the part of the universities, which reserve for themselves the right to grant admission as they deem appropriate. (This information has been described in greater detail in the discussion of Objective 1.1.a. Facet 3.)

EVALUATION

In general, the administration and handling of the item bank seems to obey clear criteria and safety protocols that provide reliability to the confidentiality of the items before their application.

However, the documentation that describes the process for item entry into the bank could be much more detailed, indicating specific criteria, for example, determining the condition with which an item is recorded, and the specific criteria to change said condition in the bank system, as well as the procedure carried out in order for such condition change to take place. It is also evidenced that the documentation does not include a detailed description of the criteria for updating or cleansing from the bank in accordance with a certain periodicity, neither does it provide academic and technical criteria to have in mind for the cleansing, as well as criteria defining the profile of those who should participate in decision making on the need to cleanse and update the bank.

With respect to the safety protocols described by the documentation, it is clear that the restricted access to the physical space of the bank, along with the profile restrictions towards access to the system, fulfill an important job with respect to the preservation of item confidentiality. Nevertheless, once again, this positive finding does not mean there is not room for improvement. For example, there is no documentation with respect to the auditing procedures to be implemented periodically in order to detect possible information leak points.

Furthermore, although the tools for item *banking* that are reviewed here have been found to be adequate, the item *selection* process, i.e., the decision-making procedures for selecting items reviewed in Objective 1.1.b. Facet 2, is not adequate.

Table 6 shows a summary evaluation for the PSU authorship tools and item bank. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 6: Summary Evaluation of PSU Authorship Tools and Item Bank

Facet 5: Authorship tools and item bank	
1. Describe the processes followed for item authorship tools and item bank.	
2. Follow-up questions (in case they are necessary and applicable)	Rating
a. How are preliminary approved items converted into candidate items to be administered in piloting? • Could you share with me four essential characteristics that commonly change between preliminary items and items ready for approval?	E
b. How are the items identified by the bank system (for example, templates)? • How does the bank system identify reading passages later connecting again to reading items?	E
c. How are parameters and characteristics of items/passages allotted?	E
d. What is the item accessibility with respect to the special needs student population (NE)?	E
e. How is the item bank safety managed?	E
f. Which is the nature of the item bank management in general? • Guidelines for updating? • Guidelines for renewal? • Guidelines regarding item exposure? • When are invalidation guidelines applied (for example, DNU)?	E

RECOMMENDATIONS

1. Overall, the tools utilized for item authoring are appropriate for the task at hand. They provide the means and the secure environment required for this type of high-stakes testing development.
2. The Item Bank captures essential characteristics of items within a secure environment. However, given the permanent technological progress in information management matters, it would be useful to periodically implement internal or external auditing systems regarding bank safety control processes, in order to identify possible vulnerable points for item confidentiality, as well as

to improve efficiency in the processes of storage, consultation and atomization of the assembly processes.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2011). *Banco de ítemes.* Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.a. Facet 6. Test distribution and test taking

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for test distribution and test taking. A framework for evaluating PSU approaches for test distribution and test taking is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.19

> The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate norms where obtained. (pp. 46-47)

Standard 3.20

> The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identify with each major area in the test classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration instructions. (p. 47)

After the tests have been assembled at the Test Construction Unit (UCP) within DEMRE and this assembly has been approved by the DEMRE Committees, the PSU is taken to the printer for plate printing. At the printer, two forms are printed of the same test (Form 1 and Form 2), as a copy control mechanism. Once the sample plate has been printed, the professionals in charge of the tests review the plate and issue their approval regarding the massive printing of the instrument. They deliver it to the Head of the Logistics Unit of DEMRE, who thereafter is totally responsible for the custody of the printing and later distribution process.

In order to carry out its work, the Logistics Unit receives, besides the sample delivered by the UCP, a distribution template, which contains information regarding the number of classrooms, the number of students, how many units of Form 1 and Form 2 are needed, as well as the reserve booklet percentage distributed for students assigned extemporaneously or for the replacement of booklets in bad printing conditions. This information is received by the Information Technologies Unit of DEMRE.

The construction of the distribution template is based upon the information obtained from the registration process on the part of the PSU applicants, during which basic information such as the location which each student has chosen for taking the test is registered. When students with special needs register, they must request in their application that the university investigate their need for being provided with special conditions (for example, the use of enlarged tests or larger size tests, in the case of slight visual disabilities).

The printing process is contracted with a private company, which allows for inspection by representatives of DEMRE to ensure the transparency of the printing process. A team formed by inspectors goes to the printing press to supervise various activities, as a security measure to prevent leaks of the test by members of the printing company's operations staff.

Some measures of security are established where the printing is taking place, such as the restriction on the presence of mobile phones on the part of the personnel participating in the printing process. There also are prohibitions concerning pencils, cameras or any device that could enable a person to reproduce any question found on the test.

As an additional security measure, the printing company is asked to deliver in DVD format all of the camera recordings registering the printing process. Additionally, upon finalizing the printing process, particular members of the Logistics Unit are responsible for eliminating the possibility of access to the trash issuing from the printing process, including the plates, blankets and inking rollers used during the printing.

Written records are kept on each material printed in order to control the integrity of printed materials both before and after their distribution. The printing company carries out the packaging of the printed material following an order of tests and forms previously provided by DEMRE. The purpose of this order is to randomize the test form distribution within each classroom. For each registered student a bag is packed with its respective test booklet and answer sheet. The bags with test booklets are packed into boxes. Control is kept over the number of large boxes and small boxes delivered to each school.

The distribution of materials to the application sites takes place in accordance with distribution templates indicating the number of classrooms in each school and the number of booklets of each form that each classroom shall have (for security purposes each classroom receives test form 1 and test form 2), which has to coincide with the total number of students taking the test. This information is drawn from the databases of registered students. Every school receives an additional percentage of booklets to replace those that may have happened to be illegible or that present some printing problem.

The distribution templates are used to control the distribution of test materials as well as their return after test administration. At the classroom or hall level, the number of tests submitted is detailed, indicating the consecutive page numbers in which every delivered package begins and ends. Each package is marked with the classroom number where it must be delivered inside the respective school.

The test packages are separated by area, in a way such as to prevent different tests from being packaged together, with the exception of schools where there is only one classroom, in which case the tests from the different areas are packaged into a single box.

Besides test distribution, the process also includes the distribution of administration materials such as pencils and erasers, as well as an unassembled box in which the students' completed answer sheets are to be placed. Once that additional box is assembled and filled, it is closed not with a seal, but with tape.

Other printed materials, such as manuals for test administration, are delivered directly to the test administrators when they are hired rather than during the distribution of test materials.

Police stations are available for the delivery and storage of test materials, where DEMRE delegates go to check out and return the test materials before and after test administration. If the test administration site is close to the police station, the policemen place a label on the boxes to certify that the materials for which they are responsible are, in fact, being received by authorized test administrators of the schools. When the test administration sites are a significant distance from the police stations, the police escort the vehicle transporting the test materials all the way to its final destination.

In total, during the test administration period between 160 and 200 trucks may be used for the distribution of the tests to all the sites. In cases of sites with no road access, there is airfreight distribution available, and the material is escorted by a DEMRE delegate (not a policeman).

To collect the materials after the test administration, the logistics are reversed. Once the test is administered, the tests are taken from the test administration site to the respective police station where trucks will be arriving, in accordance with a prescribed route, at different police stations and recording the respective return of materials. Once all of the test material has been returned to the gathering site, the booklet and answer sheet count begins, which must correspond to the number of booklets and sheets forwarded. If this is so, the booklets are sent to the designated storage site.

It must be pointed out that the test administration takes place simultaneously over all of the Chilean territory during the months of November or December, after the students who are eligible for the test (4th grade of high school) have completed their school year. It takes place on Monday and Tuesday of the chosen week to ensure availability of transportation and ease of access to the application sites.

The preparation of the delegates and technical coordinators in charge of the test administration takes place two weeks before the test administration. The remaining test administration participants are prepared the day before the test administration.

When losses of test documents have occurred before the test administration, it is because they have been wrongly classified according to their unique sequential number as they usually appear at DEMRE. In the interviews with DEMRE, their staff noted that this situation is very rare and there has not been an instance when a loss or a wrongfully placed booklet had not been found. The process of sequentially numbering booklets is filmed and a record is created. Then, when the booklets are boxed, a list indicating the sequential numbers contained in each box is created.

When the loss of a booklet occurs after the test administration, DEMRE has processes in place to track the lost booklet through its sequential number, to know the exact position where that booklet was placed in the test center facility and to identify the student that worked with it. Instances such as these are reported to the police who, with the information provided by DEMRE, can track down the student to his/her own home. These instances have occurred rarely, and, when they do, the booklets have generally been found.

There was one instance in 2004 in which three Science booklets were lost after the test administration. Those booklets were connected to three students who did not arrive to take the tests. A lawsuit was placed on the local test administrators who happened to work at a pre-university training facility, but nothing was ever proved.

EVALUATION

It is acknowledged that adequate safety protocols have been established for handling the examination materials during the printing process, and that the same may be guarantors of the non disclosure, total or partial, of the test material before an application. However, the documentation does not provide evidence for, with the same precision level, the existence of protocols for the control of material in its distribution to the application sites. Even if the participation of police stations is described in this process and that fact imprints rigorousness to the custody of the material, there is no explicit description of that which takes place with the test material once it is at the application site. For example, how does the custody of the materials of people who do not attend or that do not show up at the application date take place and by whom? Also, there is no description with respect to what protocol must be followed by the delegates who are custodians of the test materials in locations away from the capital city, in far away sites that must be reached by airplane, according to that which the documentation describes. That is, the doubt remains whether the safety level of the material may differ in the capital city and the rest of the regions.

With respect to the information gathering process in the capital city, the documentation suggests that the possibility exists of requesting copies of the delivery records in the police stations, for verification purposes, should any doubt arise. It is recommended that the collection process include, as part of the procedure, the gathering of the police station records and a comparison of the total tests rendered and collected at each test reception point. The reviewed documentation does not clearly establish the differences (should there be any) in the procedures followed for pilot tests and operational tests. It is important for the processes to be documented in detail so as to ensure follow up and control actions, as well as to duly orient the carrying out of equivalent procedures, in order to prevent the risk of affecting the validity of the assessment process due to undesirable external variables. Finally, there is no evidence in the documentation of the performance of studies or controls in support of the fact that the time available to answer the test is sufficient for the target population. This type of study is important to verify that the administration conditions themselves are not a factor affecting the validity of the assessment process.

With respect to the contingency plans for lost or misplaced test booklets, we found this process to be at the level of security required for this type of high-stakes examination. This process is comparable to those found internationally for high-stakes examinations. For example, in the United States, the delivery system of test booklets utilizes a unique identification number for each booklet for tracking purposes. This numbering process allows for reconciliation of test booklets shipped and returned.

Table 7 shows a summary evaluation for PSU test distribution and test taking (for compliance with RFP). The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 7: Summary of Test Distribution and Test Taking Process

Facet 6: Test distribution and test taking (for compliance with RFP)	
1. Describe the logistics processes followed for test distribution. Describe the test taking process.	
2. Questions for follow up (in case they are necessary and if applicable)	Rating
a. How are security issues handled? • How do you handle retrieval issues? • What type of contingency plan do you work with? • Who are involved on the receiving end at the test taking location? • What verification balance measures are taking place?	F
b. Describe the packaging process. How does the randomizing of test forms occur with respect to your operational administration?	E
c. Are there differences between the packaging and distribution plans with respect to the field test and operational forms? • Describe more about the similarities and differences between those processes.	D*
d. What are the qualifications for the test administrators? • What type of training do they receive?	E

• What type of process verification is available for the assessment of the behavior of the test administrators? I have not heard about administration manuals. • How are contingencies resolved during the test administrations? What contingency plans are available to deal with security breaches, bad weather, student health?	
e. Do you agree that the PSU is more of a power test than a speed test? • Why do you say that? • How has the research that has taken place informed the decisions taken regarding the length of the test and the time allotted for answering? • What student subpopulation participated in this research?	D*
f. In what sequence are the PSU subtests administered? • Are breaks allowed between the administrations? • Do all administrations take place in the same day? • What procedures are followed in order to minimize cheating / copying during the test administrations?	E

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. We recommend documentation for the packaging and distribution plans for the field test and operational forms.
2. We recommend clearer protocols for quality control: identity verification, copy control, and the management of chance events (crisis, illness, bad weather, etc.).
3. We recommend cataloging departures from standard administration process so that DEMRE is better equipped when facing circumstances that lead to such departures. We recommend using such a catalog to evaluate test administration process and provide training to staff participating in such processes. The professional development may allow participants in the administration process (e.g., location heads or delegates) to use their experience as PSU administrators and coordinators to report their suggestions to DEMRE on how the test administration process could be improved in the future.
4. Carrying out studies to discard the effect of time and other variables of the application (letter font, test booklet layout, instructions issued to the students, physical conditions of the locations, etc.) which may affect test performance on the part of the students.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

PSU. (2011). *Criterios de selección de locales y de personal de aplicación de pruebas oficiales PSU*. Santiago: Universidad de Chile.

Objective 1.1.a. Facet 7. Test scoring

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for test scoring. A framework for evaluating PSU approaches for test scoring is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.22

> Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if the test can be scored locally. (p. 47)

GENERAL DESCRIPTION

After the examination, the answer sheets are scanned by means of optical reading machines. The answer sheets arrive in fully identified boxes. Each box is passed by two optical readers having a different calibration with regards to the marking strength: one of them perceives weak markings, and the other one is calibrated normally. In this way, the reading of the answer sheets undergoes quality control.

It is possible to have discrepancies in the scanning of one same sheet by two machines (if there are two markings and one is very weak, it will only be captured by the machine detecting weak markings). These differences are looked into by the team called discrepancy correctors. Discrepancies found in sheets of one same box are reported in printed form identifying the sheet and the discrepancy in it.

The correction of the discrepancy is carried out by a person, a judge, who decides which is the valid mark or answer. In reality it is teams of people which consult with each other facing any doubt. This correction takes place in digital form, that is, over a file that has the answers of a sheet with the discrepancies outlined.

The judge verifies the answer sheets and performs the correction only if pertinent. For example, if the scanning generates a double answer to one question in particular and the judge, verifying the sheet finds that the second marking corresponds to a smudge (one answer is erased but the graphite from that erasure passes to another option), in that case the judge corrects, leaving the most visible marking. However, if what is happening is a double marking, the scanning marks them as a double marking and the judge leaves this marking, which is an asterisk.

When the grading takes place, the asterisks are counted as wrong answers.

In the last two assessment cycles, discrepancy reports have been generated that have included information concerning patterns of discrepancies that originated at certain administration sites. These reports provided recommendations for how these incidents can be avoided if more care is taken in the future.

Even though it is possible to do a follow up of those cases where discrepancies occur, this is not consistently done. There are also no controls over the administration

coordinators regarding the verification of the answer sheets in each session; such verification happens sporadically and only through the initiative of administration coordinators, since there are no documented procedures with respect to any verification of answer sheets.

Though not a common occurrence, some test takers make multiple markings because they have not been taught the correct manner of answering. When these multiple markings are found, the test takers are informed so that this manner of answering will not lead to low scores in the future.

In addition to the scanning provided by the two differently calibrated optical reading machines, there is also a manual verification of 1% to 2% of the answer sheets. This manual quality control procedure is carried out by the information technologies department, which is responsible for scanning and the production of databases.

The scanning file contains the following responses: a, b, c, d, e, the asterisk for multiple markings, and "O" for omissions. At the information technologies department, these characters are transformed into number data: "0" for wrong answers (including multi markings), "1" for correct answers and "9" for omissions. Once this database is fully prepared, classical test theory is used to analyze the quality of the test items. After this analysis, the committee may decide to eliminate from the final grade some questions that have poor measurement characteristics.

Once the final approved set of questions is received, the Information Technology department grades the answer sheets considering only that revised set of approved questions. There is no documentation of the scanning.

After the scanning is completed, a manual verification occurs, which involves the inspection of answer sheets chosen randomly from a representative sample of the materials.

DEMRE employs a correction-for-guessing formula to adjust raw scores when calculating final scores for each student, which includes subtracting one-fourth of the wrong answers from the total number of correct answers. Note that leaving a question unanswered is scored as zero. This procedure enables the student to take informed risks in order to obtain a particular score, since one correct answer is discounted for every four wrong ones. This procedure may generate important differences in the grading of those assessed with respect to the "knowledge" they exhibited in the test.

EVALUATION

It is acknowledged that the scoring and scanning process responds to systematic procedures with control points that endow the same with reliability. Even though the documentation of the mechanisms implemented for multiple marking and omission is clear, the documents could be studied as objects that provide information for the feedback of the complete process, e.g., identifying the causes of these phenomena, how they take place through sub populations through type of school, region, etc. By means of which enriching information could be obtained in order to improve the quality and control in other moments of the assessment process, from the test design to its administration.

It is a fact that inappropriate behavior, such as copying, introduces systematic errors unrelated to the intended test construct and invalidates test score meaning and use.

The procedures to control this type of behavior must be standardized, timely and relevant, whether applied before, during or after application of the tests. Furthermore, we recommend that the consequences for individuals violating the rules should be described and be aimed at reducing the probability of occurrence of such behaviors.

The prevention of copying by students entails monitoring mechanisms during administration of the tests, but can also include the use of specialized software for the detection of copying, or through estimates made with psychometric analysis. Although the estimation procedures that apply copying after the administration do not provide direct evidence of copying, they can guide the debugging of databases for the parameter estimation process of the items or related test score. In addition to data mining efforts, comparison of performance of test re-takers and erasure analyses are useful ways to gather additional pieces of information to documentation of test takers unethical behaviors. Finally, depending on the prevalence and severity of test takers' and test administrators' unethical behaviors, impromptu fiscal audits can be tactically articulated and allocated to the program. All efforts that have been made to address unethical behaviors during operational administrations of the PSU tests should be also made part of the pilot test administrations.

The evaluators recognized serious problems in the use of correction-for-guessing (or formula scoring) adopted for the PSU. Because of the essential role such scores play in reporting of PSU test scores, the international evaluation team regards its use as inadequate because it challenges the validity of PSU scores and PSU field test administration results. For a more detailed evaluation of this policy, see the discussion related to Objective 1.h. Facet 1: Types of scales, standardized scores and calculation procedures.

Table 8 shows a summary evaluation for PSU test scoring (for compliance with RFP). The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 8: Summary Evaluation for PSU Test Scoring

Facet 7: Test scores (for compliance with RFP)	
1. Describe the scanning and scoring processes.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Could you describe the scanning process? • What type of quality control process do you have concerning scanning? • How do you handle double markings on answer sheets? • How do you handle erasures? • What parameters do you use to inform on the quality of the scanning process and the results? • Do you develop a specifications document on score scanning/grading? o Who participate in the development of the specifications of the score scanning / grading? Why is it not important to count with a specifications document on score scanning / grading??	G
b. Where do the alpha numeric character conversions take place (for example, A, B, C, D, 1, 2, 9, missing ones)? • What reasonability verifications do you carry out over the set of scanned data before converting the raw answers into numbered answers? • What parameters do you use to inform on the quality of the graded scores data file? • What type of quality control process do you involve in the process for conversion of alpha numeric answers into numbered ones? • I have heard nothing about the use of a process (for example, case simulation) for verifying the test grading algorithms. Do you use something else? What type of assessment criteria de you use?	F
c. How are decisions made concerning the use of correction for guessing? • Explain about the empirical research you have carried out on correction due to guessing and how the results have supported the decision to adopt it. o Do you involve correction due to guessing in pilot item analysis? o And what about the operational analysis? • How do you see the consequences of retaining or	A (3.22)

extracting the correction due to guessing from the process? • What type of analysis do you carry out with respect to the questions avoided? • Have you encountered conditions where the correction due to guessing and the speed in the test were confused?	
d. Do you have ongoing processes to perform erasure analysis? • What do you do with erasure data? • Do you have statistical procedures for filtering due to cheating? • What do you do with those results?	D*

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. Even though the scanning process is thorough, the technical processes followed have not been documented nor are there any reports concerning issues that arise during each administration. We recommend that these technical processes be documented in writing and that annual reports, which record the most recent scanning issues and their resolution, be produced.
2. To date, the scanning process that DEMRE has instituted involves both mechanical and manual inspections of multiple markings. The fact that the process involves two levels of resolution on scanning plus a manual check is commendable because it reduces sources of unrelated variance. Nevertheless, though this information is primarily used for the resolution of individual scored item responses, we recommend the use of these erasure analyses at an aggregate level to further support the integrity of the test administration process. Because of the high salience of the PSU, we also recommend further analysis of potential threats to the integrity of test scores arising from unethical behavior (e.g., answer copying, etc.).
3. In general terms, no studies pointed at supporting the decisions to adjust the raw scores by correcting for guessing. We recommend implementing prospective research studies to document decisions on the use of various studies to support the decisions made. The international evaluation team also recommends a series of retrospective studies to evaluate any potential effects on PSU test scores and item banking field test statistics, for example, of the decisions made in the past due to use of formula scoring.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (25 de Agosto de 2011). *Fórmulas y consideraciones: Aprende a calcular los puntajes*. PSU en el Mercurio. Documento No. 8. http://www.demre.cl/text/publicaciones2012/agosto/publicacion11(25082011).pdf

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.b. Quality standards of question pretesting

The evaluation team developed and performed interviews with relevant stakeholders from DEMRE on March 21 of 2012. The interview process took a total of two hours from 9:30 to 11:30. The purpose of the interviews was to gain deeper understanding on the:

- Design of pilot studies (e.g., specifications, guidelines and criteria) (Facet 1)
- Decision-making process and criteria for selecting items to be piloted (Facet 2)
- Decision-making process and criteria for surveying item bank in preparation to piloting items (Facet 3)
- Process to review performance of piloted items (Facet 4)

All the interviews were performed within DEMRE offices following an agreed-upon schedule for the visit. The interviews covered the four facets and their elements as agreed upon with the TC during the goal clarification meeting in Santiago, Chile, in January 2012.

The following DEMRE staff participated in the interviews:

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees.

The following subsections contain the results of the evaluation for Objective 1.1.b., Facets 1-4.

Objective 1.1.b. Facet 1. Pilot items — Pilot design

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for pilot design. A framework for evaluating PSU approaches for pilot design is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. (p. 44)

Standard 3.8

> When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended. (p. 44)

GENERAL DESCRIPTION

Item piloting or pre-testing involves the administration of previously untested items to a sample of the students who are representative of the students who will ultimately take the PSU. The purpose of item piloting is to obtain data that are "equal" or very similar to those which could be obtained during the operational administration of the PSU. In this way, the measurement characteristics of the items may be known beforehand, and questions that produce results (scales of results) with the desired characteristics may be used during the construction of the PSU.

To verify the functional quality of the items, they are administered to a sample of students possessing characteristics similar to those of the final population.

Because the students' participation in the pre-testing is voluntary, there is an expectation that 20% of the requested sample of test takers may not participate in it.

The sampling plan selects persons with certain characteristics of the students: their region, their curricular branch, their type of school or "dependency" (i.e., *la dependencia*, e.g., municipal, private subsidized, and private paid), and their gender. (The socio-economic level of the student is not included because DEMRE has concerns about the quality of the information that has been collected for this characteristic.)

Student samples are chosen independently for piloting each PSU test. According to DEMRE's documentation, 1500 students per test form constitute the selected sample for the pilot. This quantity is usually reduced up to 20% due to lack of participation. However, DEMRE believes that the number of students actually used is sufficient to ensure the quality of the statistical analysis of IRT and DIF, i.e., it does not increase the sampling error calculated.

The sample selection takes place through stages: first, the region of the country is chosen, then the county in the region, then the establishment. In the revised documentation for the item piloting, it is not clear if all the students at each chosen establishment are sampled or only some of them.

The calculations to determine the sample are made with the optimum allocation procedure, which takes into consideration the variables indicated above. The information on the number of students that took the PSU the previous year, the average, and the standard deviation are all taken into account. The Test Construction Unit decides in which counties the piloting shall take place, based upon the information obtained for the sampling plan. The revised documentation is unclear about the criteria used for this decision.

In addition to the sample selection, the other process concerns which and how many items to pre-test. In this way, the Item Bank is supplied with piloted items that have the characteristics needed for inclusion in the final test forms that the students will ultimately take.

Our review found no rationale documenting the definition of how many and which items to pilot in each one subject area of the PSU. However, the Item Bank is analyzed annually to determine the current needs, given the results of the previous year's pre-testing, i.e., the number of items that were left over after the previous year's test forms had been constructed.

EVALUATION

The item piloting process works with student samples selected taking into account socio demographic variables that suggest the formation of representative samples. However, because the students attend the pilot application voluntarily and apparently with no particular impact for them, it may be predicted that the motivation level is different with respect to the target population of the test, which would explain the differences in the indices of classical test theory presented in Objective 1.1.f. This may have negative effects in the estimations of the item parameters, meanwhile phenomena such as omission or guessing may increase sensibly. This should be analyzed and documented. Carrying out studies analyzing the variation ranges of the statistical indicators between piloting and the operational application is fundamental in the orientation of the decision making on pertinent strategies for assuring that the item piloting really responds to the need of calibrating item parameters and for recognizing the quality of same, before the operational application. In general, the documentation of this process requires adjustment in order to provide it with greater precision levels with respect to, for example, the description of criteria for the definition of item needs in the bank with a view to planning a pilot application and the criteria for defining the population sizes for each test, among others.

The procedure for selecting the sample adheres to accepted sampling criteria, taking into consideration aspects of the strata that are particularly important for PSU testing (dependency, type of curriculum, etc.). However, this straightforward design may be affected in practice by the non-participation of sampled students whose involvement in the pilot is strictly voluntary. Acknowledging the potential for non-participation does not help solve how to achieve greater participation. It may be the case that instead of discounting the potential non-participation rates, the rate should be built on top of the target sample sizes in such a way that when non-participation peaks the end results are closer to the intended target *for each and every subject test form piloted*.

Nevertheless, increasing the number of students taking the pilot will not, by itself, solve the problem that the samples may not be representative. The high rate of non-participation may be systematic rather than at random, thus threatening the intended representation of pilot samples over relevant socio-demographic variables.

Taking into account what was found in other objectives (such as Objective 1.1.f., for example), it is clear that the pilot process incorporates procedures performed in an orderly manner. However, there is a lack of a general purpose driving the piloting of the items and psychometric expectations on pilot results. If the purpose of the pilot testing is to gather item data to be further analyzed by groups of reviewers in item data review sessions, the procedures should clearly state boundaries of psychometric performance expected for the items and nature and representation of review panels. If the expectation is to estimate pilot performance of items to inform test construction without involvement of data review meetings, something which is necessary for a high stake test such as the PSU, there is evidence that this purpose is not fulfilled because the data indicate drastic changes in item properties between pilot and operational administrations.

Table 9 shows a summary evaluation for the PSU pilot design. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 9: Summary of Evaluation of PSU Pilot Design Processes

Facet 1: Pilot design	
1. Describe the specifications, guidelines and criteria used in the item piloting process.	
2. Questions for follow up (in case they are necessary and if applicable)	Rating
a. Describe the specifications with respect to data collection and analysis for the PSU pilot study. • Which is the process for defining the needs with respect to pilot studies and research? • Who writes (or determines) the needs and which are their requirements? • How are the specifications updated and how often do the updates take place? Why or why not? • What is the sample unit (student or school)? • Which is the process you follow to achieve your objectives regarding pilot studies? Do you base yourself on a registration file to allocate the pilot forms? • Do you randomize the pilot test forms within the testing location? • What incentives are provided for the pilot sampling units? • How many pilot administrations take place in a year? When? • In case of more than one pilot administration, do you base yourself on the same sample? Do you extract a different sample? o What verifications are performed on sample variability? o Does the sample composition remain the same in administrations within the same year? ▪ What about in case of administrations between years? o What information collection focus is followed?	A (3.7; 3.8)
b. Is the pilot design appropriate for the PSU target population (for example, demographic representation, region, SES, curricular branch, type of school)? • What types of considerations are given to the changes in the target population taking the operational PSU test? • How are the foreseen and unforeseen consequences assessed? • No mention has been made of research in support of the decisions. Could you provide me with a rationale?	C (3.8)
c. What is the process for predicting the numbers of items and the number of reading passages to be administered? ▪ Who participates from the prediction decisions? ▪ Who writes / approves the required documentation?	A (3.7; 3.8)
d. What is the rationale behind the numbers of piloted items per administrative cycle?	A (3.7; 3.8)

▪ No mention has been made of about taking into account the rejection rates due to an inadequate item performance. Could you please clarify this?	
e. Do you pilot all of the PSU items or do you choose one type instead of another? What is the process with respect to items based upon passages?	E

RECOMMENDATIONS

1. It is necessary to establish a sound purpose for the pilot. First, rethink the whole process of the pilot by defining the goals and the use and the procedures to be carried out according to this definition. For example, develop sample size quotas that take into account expected rates of non-participation—and that those rates might be different for different subjects—so that the goal of 1500 participants is uniformly met. Furthermore, analyze the impact of non-participation rates on representation of major socio-demographic variables. Next, find socially acceptable ways to increase students' motivation to give their maximum performance on pilot administrations. Finally, identify clearly the quality of the items expected and obtain preliminary values of the parameters that are consistent with the final administration. From the results of the Objectives 1.1.f. and 1.1.g., the pilot administration has little value beyond the marginal analysis of the quality of the items; hence, the low rating in some respects.
2. We recommend that additional documentation be provided for the following areas: the rationale behind the pilot design; data collection and analysis for the PSU pilot study; and the process for predicting the numbers of items and the number of reading passages to be administered.
3. Although the statistical criteria for the sampling plan for the pre-test have been documented, we recommend, in particular, better documentation of the criteria for the choice of counties (*comuna*) in which the piloting is to occur.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2010). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales.* (Admisión 2012). Santiago: Universidad de Chile.

DEMRE. (2011a). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2011b). *Procedimientos para determinar la muestra para el pre-test*. Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.b. Facet 2. Pilot items — Item selection

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for selecting items. A framework for evaluating PSU approaches for selecting items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. (p. 44)

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures and evidence of model fit should be documented. (p. 45)

GENERAL DESCRIPTION

The members of the UCP are those responsible for selecting the items to be piloted to repopulate the item bank. Piloting includes items recently constructed and approved by the Committee of each respective field or those that have already been piloted but which required important modifications.

The committee of each respective PSU test approves by consensus which items are to be piloted. The statement of the item, the answer options, and the correct answers are verified. The UCP committees use the item construction rules and the expected item difficulties to guide their selection of items to pilot. There is no reference to test construction tools aiding the construction of the forms.

Additionally, expertise of the participants also plays a role in the decisions over and above the aforementioned criteria for decision making. The use of the expertise of the participants to make judgments about the items has not been documented and no reference is made regarding the documentation thereof, if it exists, in the reviewed documents or in the interviews.

Once the items approved for piloting have been entered into the Item Bank, the assembly of the pilot tests proceeds. Neither the criteria for this assembly nor the criteria for the number of items nor the organization of same by curricular subjects or criteria are clearly described in technical documentation. For example, it is unknown the role of item response theory estimates (i.e., the difficulty and discrimination

parameters that are used in the item analysis by DEMRE) on test construction. It is also does not provide sufficient detail concerning the role that pilot items plays across pilot forms. That is, whether common items have been embedded across pilot forms has not been demonstrated from the documentation or the interviews.

EVALUATION

Although the tools for item *banking* that were reviewed in Objective 1.1.a. Facet 5 have been found to be adequate, it does not follow that the item *selection* processes, i.e., the decision-making procedures for selecting items, are adequate.

The documentation concerning this facet is insufficient and during the interviews additional information was not forthcoming. The decisions with respect to the item selection for piloting seem to address a single objective, which is to close the gap between current and expected item counts in the bank. However, according to Standard 3.9, this process also needs to reflect the psychometric characteristics of the items expected from piloting. The description of this process does not permit the identification of additional criteria for selecting items for the pilot. These additional considerations include exploring the psychometric effect of different item formats or the statistical behavior of items based on their location within the test booklet, among others.

Table 10 shows a summary evaluation for PSU pilot item selection. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment

standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 10: Summary Evaluation of PSU Pilot Item Selection

Facet 2: Pilot item selection	
1. Describe the processes and criteria used regarding the selection of items to be piloted.	
2. Questions for follow up (in case they are necessary and if applicable)	Rating
a. What process is used in the selection of items for testing? • Does the selection involve a filter carried out by the committees? • Who forms part of those committees? • How soon during the process do the meetings take place? • How long does the process remain the same with respect to the Language and Communication items?	A (3.7; 3.9)
b. Which are the content, statistical and population criteria with respect to the selection of items to be piloted? • Which is the rationality behind the criteria adopted? There has been no mention of anything about the rationale behind the population criteria. Could you please clarify this? • How are the criteria, who participates and the frequency at which the updates take place updated? • How is it that the predicted use of PSU scores (NRT vs. CRT) are to be considered in the criteria? • How is it that the curricular branch differentiation is considered? • How is it that the type of school (for example, private, subsidized, municipal) is considered? • Are the region and SES considered throughout the process?	A (3.7; 3.9)
c. Which is the framework for the development of pilot forms? • Since they are forms alone by themselves, how do you go about constructing them? • Do you involve an objective in particular when developing pilot forms? • Would it be fair to state that the pilot forms look like the PSU operational forms regarding their content? Within a pilot form, what roles could the pilot form perform? • Are the items repeated throughout the forms? • As far as the reading passages, do they repeat themselves? • What prevents the use of pilot items incorporated into the operational administration of the PSU?	A (3.7; 3.9)
d. What is the proportion / size of the sample / pilot item?	

• What is the proportion for passage based items? • Has it been historically proven that this proportion is the optimum? • How well does the proportion capture the student diversity in terms of demography, region, curricular branch, socio-economic condition?	A (3.7; 3.9)

RECOMMENDATIONS

1. We recommend that documentation of the criteria for choosing pilot items be supplemented with a systematic articulation of the reasons for item selection that are determined by the expertise of the participants. In accordance with Standard 3.7, the documentation of these processes would ensure the repeatability of the same even when the expert groups involved in the development vary, thereby increasing the reliability of the process implemented. In this sense, we recommend that greater details be provided, including a statistical rationale, with respect to the placement of common pilot items embedded across more than one pilot form.
2. The recommendation is for the planning of the pilot applications to include clear and intended objectives towards the verification of aspects such as the psychometric effect of using the same items with different item group blocks, or on the effect of the change in position of an item in different booklets, etc. These studies must be documented and should give feedback to test design.
3. We recommend documenting the purpose of the pilot and the rationale for determining what items are needed according to standardized specifications. The documentation for the pilot must also contain the requirements for sampling and analysis of the items after administration and the elements of CTT or IRT that are considered relevant to the design of the pilot.
4. Perform analysis of the documentation for each cycle of piloting and rate its compliance with the procedures described in the previous section. This analysis should be done after the pilot administration and should be well documented. Checklists can be devised to document fulfillment of the pilot specifications, processes and their stages. Overall a system of quality checks should be put in place to monitor the quality and usefulness of the pilot components and outcomes.

BIBLIOGRAPHY

DEMRE. (2010). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales 2011.* (Admisión 2012). Santiago: Universidad de Chile.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2011). *Procedimientos para determinar la muestra para el pre-test*. Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.b. Facet 3. Pilot items — Item bank analysis in preparation for piloting items

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for preparing for pilot items. A framework for evaluating PSU approaches for preparing for pilot items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. (p. 44)

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures and evidence of model fit should be documented. (p. 45)

ETS Standards for Quality and Fairness (ETS, 2002)

Standard 2.3

> Establish and document procedures to maintain the technical quality, utility, and fairness of the product or service. Once a program establishes the technical quality, utility, and fairness of a product or service, the program must carry out procedures, such as periodic reviews, to ensure that the suitability is maintained. In general, five years should be the longest interval between reviews. (p. 12)

GENERAL DESCRIPTION

In general terms, an administrative process takes place each year, which includes the revision of items for piloting and which begins with the hiring of item writers. The number of item writers hired is directly related to the quantity of items needed as determined by an audit of the Item Bank. To replenish the deficits in the Item Bank, item writers will attend to the CMOs, the cognitive skills, percentage distribution by thematic axis, estimated difficulty, and classes of items found in particular test sections.

These constructors are selected from across the nation, where basic education, high school and university teachers may participate, who must fulfill a series of requirements already mentioned in another objective.

At the beginning of their activity, the constructors must sign a confidentiality affidavit concerning the material they will produce.

Each constructor is provided with a road map sheet detailing the items that he or she must construct each week and in total. The constructed items of a particular area are reviewed weekly by the respective committee, which will decide about their inclusion into the Bank.

To guarantee equality in item construction, the constructors are trained in the criteria used for item construction, especially in those aspects of items that are taken into account by the committees prior to their approval.

After the training workshop, the constructors, with their road map sheets, initiate the process for the construction of items by using their personal computers to log into the Safe Question program. Once delivered via the Safe Question program, the constructed questions are then put on a USB memory stick at DEMRE for revision by the respective committee. If the items are approved, they are stored on another USB memory sick until the moment when the committee member responsible for them loads them into the Item Bank platform.

Approved items then move to the piloting stage, which is automatically initiated when the responsible person on the committee chooses items for the assembly of the pilot forms. After being piloted, the item analysis statistics are generated and reviewed by the commission. If the item is approved during this data review, it goes (along with its item statistics) into a new condition in the Bank, which indicates it has already been piloted and that it may be used in assembling the definitive test. If, on the contrary, the item is rejected, this rejection is recorded in the Item Bank system, which blocks it from being used in other processes.

All of the aforementioned is registered in the system, which enables the committee to track the actions performed with the items along with the piloting results.

Those managing the Bank are the committee heads or the head of the Item Construction Unit. They are persons with ample experience (more than five years in DEMRE) and with the sufficient academic backgrounds (master's or doctorate degrees).

EVALUATION

According to the technical documentation, the process for item banking needs analysis with respect to the preparation of item construction and the piloting that takes place periodically at the beginning of each year. This existing review seems to focus itself fundamentally on the lack of specifications matrix coverage of each test, which constitutes a valid and important criterion. However, it leaves aside the possibility of designing piloting on a scientific basis to study the psychometric effects on the items, item length, booklet editing, etc., in order to enrich assessment process decision-making with them, from the design to administration. The item bank review is carried out independently by personnel responsible for the test, and the documentation does not indicate that standardized criteria are followed to perform such reviews.

Table 11 shows a summary evaluation for PSU item bank analysis in preparation for piloting items. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 11: Summary Evaluation of PSU Item Bank Analysis in the Preparation for Piloting Items

Facet 3: Item bank analysis in preparation for piloting items	
1. Describe the processes and criteria followed in item bank surveys when preparing the items for piloting.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. How often is the PSU item bank analyzed for purposes regarding the needs for the pilot testing of more items? • Are there specifications placed in position for directing the search and report results? • What criteria are involved in the bank review? • How early during the admissions process does the item bank survey take place?	E
b. What are the qualifications with respect to the personnel managing the item bank? • Is the item bank process documented and utilized during the training efforts?	F

• What is the process concerning bank update after the review of the pilot tests take place? What pilot item characteristics are updated in the item bank? • What happens with the rejected pilot items?	
c. Which is the review process for the item bank survey results? • Are there process verifications available for the review and what do they point out?	D*
d. What is the communications process for the survey results among the functional groups? • Is there a formal document where the results are incorporated? (for example, the construction process for pilot tests)	D*

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. We recommend documenting the review process for the item bank survey and the establishment of standardized criteria guiding such processes for all tests, or, in case it is necessary, the justification in order for that process to take place differently for each test. Having manuals to provide the rationale for the pilot analyses ensures that decisions made about aspects of piloting do not neglect the statistical criteria for selecting items.
2. We recommend documenting the plan for communicating the survey results among the functional groups. Pilot studies results should be issued in a systematic way among the teams responsible for the tests. For the piloting to be effective, it should contribute to the improvement of the design, construction, review and assembly of the test.

   These recommendations are consistent with international institutions like the Educational Testing Service Standard 2.3, which recommends documenting the quality of the instruments used. As mentioned in the above facet, it is important to document the practices that take place during piloting of the items. Furthermore, this effort should go beyond noting the insufficiency of the item bank and stating how to find items that meet the statistical criteria.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2010). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales 2011.* (Admisión 2012).

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago.

DEMRE. (2011). *Procedimientos para determinar la muestra para el pre-test*. Santiago.

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.b Facet 4. Pilot items – Review of the pilot item performance

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for reviewing pilot items. A framework for evaluating PSU approaches for reviewing pilot items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures and evidence of model fit should be documented. (p. 45)

GENERAL DESCRIPTION

After the pilot application, each committee reviews the information obtained. One part of the information is collected directly from those evaluated, who indicate what they think about the items. These statements are taken into account in making decisions about the future state of the items. Additionally, the statistical information derived from the application is produced and reviewed.

The classical test theory statistics are produced by the item bank system itself, whereas the item response theory statistics are produced by the Studies Unit using BILOG 3.11 software (Mislevy & Bock, 1998).

A series of criteria are used for determining if an item has behaved in an acceptable manner from the statistical point of view. An item is accepted for future applications or for its definitive form if the following conditions are met:

- The biserial correlation for the correct answer is > 0.25.
- It has no positive biserial correlations in the other options.
- The proportion of correct answers (p) >0.10.
- The non-response (without answer) rate is not p>0.3.
- It has a high discrimination value (parameter a) in IRT.
- It has a difficulty value within range (parameter b) in IRT.
- There is no distractor with a higher average than the correct answer.
- The p-value is in the range.
- The omissions are considered "low," when the omission rate is <50%.

The review of all the information is carried out by the whole committee of the field, which has been formed to be as demographically, professionally and geographically diversified as possible. Nevertheless, even with this manifest diversity, the intention of each committee is to arrive at a consensus regarding its decisions. The accepted items

enter into the item bank and the rejected items are used in future constructor training processes, illustrating the faults in the same which should be prevented in new items.

EVALUATION

The results of the pilot administration should be carried out as a team and not individually. This allows the inclusion, along with the statistical criteria, of some content considerations of the items or of the reports of the assessed students which could lead to the elimination of an item which eventually may fulfill all of the minimum parameters established or which may eventually not comply with one of the minimum parameters established. However, the documentation does not specify the decision route should there happen to be a conflict between the expert criterion of the technical teams reviewing the results of the pilot and the established statistical criteria. The criteria that should be considered and the order of their priority for this type of decision should be documented.

The criteria established for most item data review statistics are reasonable. They are in line with what is seen internationally; specifically, the literature and manuals for software commonly used in psychometrics and instrument evaluation. For example, the recommendation within ITEMAN software is for biserial correlations values above 0.25; this is the same value that is suggested in the technical reports of PISA and other international studies.

However, the upper limit for a reasonable omission rate (i.e., reasonable <50%) is higher than that seen in other programs in our experience. There is no rationale provided for setting such a high rate for reasonableness of omissions. Historically, the omission rate that we found when analyzing PSU pilot statistics was 25.6%. Given the great difference between the upper limit and the rate found, the current criterion would have no practical effect, which indicates to us that the upper limit is set too high.

It should be noted that the number of questions omitted by international educational assessments (such as PISA) is not as high as that reached by the PSU assessments. According to the *PISA 2006 Technical Report* (Organisation for Economic Co-Operation and Development, 2009, p. 219), the weighted average of omitted items was 5.41. In the *PISA 2009 Technical Report* (Organisation for Economic Co-Operation and Development, 2012, p. 200) the average number of omitted items, 4.64, was slighting smaller than in 2006. This suggests that a reasonable upper limit for omissions might be more in the order of 10%.

Table 12 shows a summary evaluation for PSU review of item pilot performance. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled “Rating” in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 12: Summary Evaluation of PSU Review of Item Pilot Performance

Facet 4: Review of item pilot performance	
1. Describe the processes and criteria followed in the review of Pilot Item information	
2. Questions for follow up (in case they are necessary and if applicable)	Rating
a. How often do data reviews take place?	E
b. When do data review meetings take place? • Who participates in the pilot data review meetings? • How is the pilot item information reviewed? • Who participates in the review? • Are there educator panels involved in the review? • What type of training is provided to the educators as part of the information review process?	C (3.9)
c. Are the participants in the pilot information review demographically diverse? • What is the composition of the reviewing group in terms of curricular branch, region, type of school (for example, municipal, private, subsidized), curriculum specialist, university faculty and region?	E
d. Are the participants professionally diverse? What type of professionals do you involve in the review of the pilot	E

information (high school teachers, university professors, curriculum specialists)?	
e. Are the participants geographically diverse?	E
f. What materials are used in carrying out the pilot information review? • What criteria are followed for the review of the pilot information? • How do you consider the predicted use and targeted for testing populations during the review? What process is followed to resolve statistical flags on item difficulty, item discrimination and item differential functioning? • What type of information is reviewed by panels? Would you do me the favor of showing me an example of one item card used in information review? • Does the review take place online or with pencil and paper? If it is online, could you indicate the outstanding characteristics of the review process and of the on line tool?	E
g. What is done with the results of the information review? • What happens with the rejected pilot items? • What is the process for updating the bank after the pilot testing reviews take place? • What pilot item characteristics are updated in the bank? • What are the policies on minor intrusions of the content?	E

RECOMMENDATIONS

1. We recommend additional documentation concerning the pilot test data review meetings, including information concerning the background of the participants and the training they received. It requires having manuals to guide this process to ensure that, while the expert teams are different from one application to another, the criteria for reviewing and analyzing are systematically maintained.
2. We recommend that the results of each pilot administration be used to progressively improve the information provided during the training of future item writers. The analysis of approved and unapproved items along with the statistical and psychometric information helps to qualify the processes of item development, which facilitates discussion about those items with the new item developers.
3. Although most of the statistical criteria used for item data review are consistent with international standards regarding acceptable ranges for psychometric indicators, we recommend a re-evaluation of the criterion used for reasonable omission rates. While there is no single international standard regarding omissions acceptable range, it is necessary to provide arguments to substantiate the criteria used for the rate of omissions. Following the comparison to PISA describe in our evaluation, we suggest that a reasonable upper limit for omissions might be more in the order of 10%.
4. High omission rates can have an effect on other statistical criteria as well. For example, calculating a CTT item difficulty based on little more than half of the population has implications for the accuracy of estimated parameters, the final assembly of operational test and, ultimately, the validity of the test results.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Assessment Systems Corporation. (2012). ITEMAN 4.1.: Classical item analysis. [Software]. St. Paul, MN: Author.

DEMRE. (2010). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales 2011.* (Admisión 2012). Santiago: Universidad de Chile.

DEMRE. (2011a). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile. Santiago: Universidad de Chile.

DEMRE. (2011b). *Procedimientos para determinar la muestra para el pre-test*. Santiago: Universidad de Chile.

Mislevy, R. J., & Bock, R. D. (1998). BILOG 3.11 Windows. [CD-ROM]. New York, NY: Psychology Press.

Organisation for Economic Co-Operation and Development. (2009). *PISA 2006 technical report.* OECD Publishing. Retrieved from: http://www.oecd.org/pisa/pisaproducts/pisa2006/42025182.pdf

Organisation for Economic Co-Operation and Development. (2012). *PISA 2009 technical report.* OECD Publishing. Retrieved from: http://www.oecd.org/pisa/pisaproducts/pisa2009/50036771.pdf

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.c. Criteria for question selection for the assembly of definitive tests

Processes and policies for test assembly should include both content and psychometric guidelines for item selection and sequencing. Often these guidelines operate in conflict with one another. The evaluation team developed and performed interviews with relevant stakeholders from DEMRE on March 21 of 2012. The interview process took a total of 2 hours from 11:45 to 13:45. The purpose of the interview was to gain deeper understanding on the:

- Intended and unintended uses and meaning of PSU test scores and intended population of test takers (Facet 1)
- PSU test design and specifications followed when pulling an operational form (Facet 2)
- PSU test construction specifications (Facet 3)
- PSU specification matrix orienting pulling of an operational form (Facet 4)
- Process for pulling an operational PSU form and the criteria followed for the pulling process (Facet 5)
- Process and criteria for reviewing and approving a pulled operational PSU form (Facet 6)

All the interviews were performed within DEMRE offices following an agreed-upon schedule for the visit. The interview covered the six facets and relevant elements as agreed upon with the TC during the goal clarification meeting in Santiago, Chile, in January 2012.

The following DEMRE staff participated in the interviews:

- Head of the department of test construction
- Coordinator of test construction committees
    - Mathematics
    - Language
    - History and Social Studies
    - Science

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees.

The following subsections contain the results of the evaluation for Objective 1.1.c., Facets 1–6.

Objective 1.1.c. Facet 1. Purposes of the PSU

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process concerning its purpose. A framework for evaluating PSU approaches for assessing PSU's purpose is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 1.4

> If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary. (p. 18)

Standard 3.1

> Test and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development. (p. 43)

Standard 3.2

> The purpose of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose of the test and about the relation of items to the dimensions of the domain they are intended to represent. (p. 43)

GENERAL DESCRIPTION

The PSU is the assessment instrument utilized by the universities forming part of the *Consejo de Rectores de Universidades Chilenas* (CRUCH) to select applicants for the openings offered by these universities. The single purpose of the PSU "consists in the elaboration of an application ranking" (DEMRE, 2010c, p. 6), starting from the assessment of curricular contents and intellectual skills extracted from the national curriculum elaborated by the Ministry of Education in 1998 for the four grades corresponding to high school education.

The historical antecedents to the PSU were the *Prueba de Bachillerato*, used by the Universidad de Chile for the 116 years between 1850 and 1966, and the *Prueba de Aptitud Académica* (PAA), which was used between 1967 and 2003. The PSU arises from the mandate of the CRUCH in the year 2000, in the midst of a political context underlined by the educational reform of the 1990s, the dissatisfaction with the PAA and the *Pruebas de Conocimientos Específicos* (PCE), which together had been used for entrance into the universities until that date, and the controversy unleashed around the project for the creation of the *Sistema de Ingreso a la Educación Superior* (SIES) between 2001 and 2003 (DEMRE, 2010c).

According to the technical documentation of DEMRE, the organization responsible for the design and application of the PSU, the PSU is a standardized, multiple-choice, norm-referenced assessment that is administered with pencil and paper. As such, its scores or results are expressed by means of standardized scales, which rank each

student within a distribution of a reference group, which consists of the total number of students who have taken the test during each administration. This is the basis for carrying out the student ranking which shall be used by universities to select students for admission as well as to allocate scholarships to applicants from low socioeconomic strata, within a government program that seeks to promote higher education among the economically disadvantaged.

In its design, the PSU is fundamentally an instrument that measures the knowledge and skills embedded in the curricular contents defined for high school education with the purpose of providing an indicator for the selection of applicants to higher education. As it seems to happen with these types of assessments, DEMRE acknowledges that the PSU reports may be interpreted by some part of the population, particularly by high school education institutions, as an indicator of the quality of the education of the students at this level. DEMRE publishes technical documentation through its web site as an attempt at clarifying use of PSU scores to discourage different uses of the results. It also uses mass media (e.g., *El Mercurio* newspaper) to communicate the official use of PSU results and the technical characteristics of the instrument.

Additionally, when DEMRE reports PSU results to higher education institutions, it issues additional warnings that the proper use of PSU results is only for informing admissions decisions and that other uses of the results are neither allowed nor approved by DEMRE.

Finally, when DEMRE reports PSU results to high school buildings, it does so in a way that intended uses of the scores appear with caveats. For example, DEMRE's publication, *Sistema de Información de los Resultados de las Pruebas de Admisión a la Educación Superior* (SIRPAES) states:

> [T]he results obtained by their students, in no case, can be understood as an evaluation of the quality of education provided by the educational establishment. Without prejudice to the foregoing, and circumscribed by the PSU framework, it is also true that they provide guidance for the educational establishment as the observed results to infer: (a) the ability of the student to implement the contents and cognitive skills acquired throughout their secondary education and (b) guide the educational establishment about the strengths and weaknesses of your group of students, when faced with and assimilate the teaching-learning process received. (DEMRE, 2012b)

The SIRPAES report has created a controversy on an unintended use of the PSU test scores, e.g., to assess the academic quality of high schools. The report brings limited descriptions of the intended and unintended interpretations of the PSU scores and leaves users to draw their own inferences.

EVALUATION

It is acknowledged that the PSU is a relatively new test, taking into account not only its first application date, but as well the fact that its object of assessment was completed progressively over time. For this reason, it is not possible to talk about substantial changes in the meaning and use of the test scores throughout its history. Even so, it would be expected that as of this date there would have been studies available which would detail the perception of the different test users (direct and indirect). Such studies could already be providing information for deciding how to adjust the PSU content, as well as its formal aspects (edition), its conditions for administration and the release and use of the test results.

There are additional uses of the PSU test results that are not intended, such as those when using the SIRPAES report to pass judgment on the quality of schools. As noted in our general description, DEMRE has made some attempts at providing caveats concerning the use of the report, e.g., not drawing conclusions about the quality of education provided at particular institutions from the PSU results. However, it is not enough to put a caveat on a web site; such messages must be embedded in the report itself. The caveats included in the report are insufficient for clearly communicating the intended use of the PSU results. As a result, the likelihood of misusing the PSU results, e.g., disaggregation of scores by school and comparisons with other buildings, is high.

From an international perspective, the evaluation team's experience with the dissemination of university examination results (like SIRPAES) is that such distribution is limited (beyond that to the targeted universities themselves) to the individual applicants and their high school counselors. The primary purpose of this reporting is to look forward to university admission for each student rather than to look back at the quality of the secondary education institution. In the United States, the quality of secondary education is adjudged by statewide assessments specifically designed for that purpose.

When acknowledging the difficulty of creating an assessment culture which aims to provide for the sensible use of the results, it is advisable to multiply the efforts at spreading precise information on the interpretation scopes and limitations that are possible to make with the test results to the appropriate audiences. Even if the Internet and, in general, mass communications media fulfill an important function at providing information, advances could be made in more direct means for releasing information (conferences, teleconferences, seminars with the academic community, etc.).

A practice that could become useful would be to focus on specific audiences when planning and developing ways to release PSU results. Journalists and other representatives of the communications media are important allies for the precise targeting of the interpretation of the results. It is sometimes useful to call for press conferences in which the journalists would be "trained" regarding what the test results may or may not be saying. This training could reduce the number of newspaper reports that over-generalize the results and thereby confuse readers concerning what the test results may really be saying. Furthermore, it would be useful if directors of high schools received information that oriented them as to what interpretations can be drawn from the PSU reports for their students, as well as warning them about the risks of using these reports for unintended purposes.

The diversity that exists among the users of the PSU would demand a great effort—in material and human resources—to directly and effectively communicate to them the restrictions regarding the use of the results. In fact, the PSU has the students finishing their high school education and who aim to enter higher education as target population objectives, and it is these students, along with the Universities using these results for student selection, who are the main users of the results.

Nevertheless, even the caveats provided by DEMRE (e.g., "(a) the ability of the student to implement the contents and cognitive skills acquired throughout their secondary education") need more support in the form of validity studies that would document—to a great degree—that the PSU assesses mastery of the cognitive skills currently found in the high school curriculum.

Table 13 shows a summary evaluation for the purpose(s) of the PSU test. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 13: Summary Evaluation of PSU Purpose(s) of the PSU Test

Facet 1: Purpose(s) of the PSU test	
1. Describe the intended and non intended uses and the meaning of the PSU scores and the test takers.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. What is the history of the PSU test? • Which are the three most critical changes in the meaning and use of the PSU test score and the intended population of test takers? • How has research informed with respect to said changes?	C (3.1)
b. Which are the intended uses and the meaning of the test scores (for example, placement, admissions, selection, scholarships, school assessment, responsibility, NRT, CRT)? • What is the process followed when reviewing the intended use/uses and the meaning of the test scores? • Which are the intended test taker populations? • What is the role of intended use and test score meaning review research, and of the test taker populations?	E
c. Are there any other uses and meanings with respect to the PSU test scores? • Are there unintended uses and meanings of the PSU scores? • What actions have been taken to discourage unintended uses and meanings of the test scores?	C (1.4)
d. Are there any other users of the PSU test scores? • Are there unintended users of the PSU test scores? • What actions do you take to educate unintended users of the PSU scores in discouraging the uses and meaning of the test scores?	C (1.4)

RECOMMENDATIONS

1. We recommend continuing the efforts to educate the public about the intended uses of the PSU declared in the official documentation available from DEMRE (the selection of students for admission by universities) and MINEDUC (the associated granting of scholarships to economically disadvantaged university applicants). Specifically, DEMRE and MINEDUC should plan and develop strategies for releasing information to special audiences such as education journalists and directors of high schools, indicating in a clear manner what the appropriate interpretations and the restrictions are for the use of the results. Such strategies could strengthen the validity of the entire assessment process.
2. We recommend caution when developing new reports to be shared with larger audiences (e.g., SIRPAES). At minimum, reports should be reviewed by policymakers for their unintended as well as intended consequences. Along these lines, we also recommend clearly stating disclaimers of unintended uses of these reports and launching efforts to educate the general audience regarding the intended uses of the reports. It would be commendable developing case studies to provided examples of intended and unintended uses of the reports.

3. We strongly recommend surveying the uses of test results that were never intended and proactively discouraging such uses until validity evidence is gathered to support them. For example, SIRPAES reports back scores based on cognitive skills within each subject area. Evidence should be gathered to support the reliability and validity of using these scores. In addition, because the educational training provided by secondary schools in Chile is influenced by many factors above and beyond those factors measured by the PSU, the general public should be cautioned about too readily drawing conclusions about the PSU based on the student results gathered from any particular type of school (*la dependencia*) or curricular branch (*la modalidad*).

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2010a). *Descripción técnica de la prueba de matemática.* Santiago: Universidad de Chile.

DEMRE. (2010b). *Marco teórico prueba de selección universitaria historia y ciencias sociales.* Santiago: Universidad de Chile.

DEMRE. (2010c). *Prueba de selección universitaria (PSU): Antecedentes y especificaciones técnicas.* Santiago: Universidad de Chile.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2012a). *Actualización marco teórico ciencias naturales (física).* Santiago: Universidad de Chile.

DEMRE. (2012b). *Desde 5 de abril, SIRPAES del Proceso de Admisión 2012.* Retrieved from http://www.demre.cl/noticias/not_120403_sirpaes_disponible.htm

DEMRE. (2012c). *Marco teórico de la prueba de selección universitaria (PSU) del sector ciencias naturales subsector de biología*. Santiago: Universidad de Chile.

DEMRE. (2012d). *Marco teórico prueba de lenguaje y comunicación.* Santiago: Universidad de Chile.

DEMRE. (2012e). *Marco teórico PSU-ciencias-química.* Santiago: Universidad de Chile.

Objective 1.1.c. Facet 2. PSU test design

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process to design the PSU. A framework for evaluating PSU approaches for designing the PSU is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 2.1

> For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported. (p. 31)

Standard 3.3

> The test specifications should be documented, along with their rationale and the process by which they are developed. The test specifications should define the content of the test, the proposed number of items, the items formats, the desired psychometric properties of the items and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information. (p. 43)

Standard 3.4

> The procedures used to interpret test scores, and, when appropriate, the normative or standardization samples or the criterion used should be documented. (p. 43)

Standard 3.6

> The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses, and when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences and demographic characteristics of expert judges should also be documented. (p. 44)

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subsets according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented. (p. 44)

Standard 5.7

Test users have the responsibility of protecting the security of test materials at all times. (p. 64)

GENERAL DESCRIPTION

Although the mandate for developing the PSU tests came from the CRUCH, DEMRE is responsible for the technical design and administration of the tests. The PSU has as its target population the students who have finished their high school education and who wish to enter a university within the CRUCH and affiliated universities. The proposed purpose of the test is assessing the cognitive reasoning capacities, including intellectual and problem resolution skills, in areas fundamental to the national high school curriculum. The PSU battery comprises tests in the fields of: Language and Communication, Mathematics, History and Social Sciences and Science. The Science test includes a common section and an elective section in one of the following subjects: Biology, Physics or Chemistry (DEMRE, 2010c). The difference between a common part and another elective one corresponds to the differences found among the universities' requirements with respect to the programs that they offer.

The Mathematics test and Language and Communication test are mandatory, while the Science test and the History and Social Sciences test are elective. Students are obliged to choose at least one of the elective tests. The test frameworks reference domains, defined by Fundamental Objectives (OF) and Mandated Minimal Content (CMO), which were formulated as part of Chile's national curriculum for middle schools. Over time, the PSU test frameworks came to be adjusted as the assessment system was implemented.

It must be pointed out that the national curriculum places emphasis on two curricular branches: Scientific-Humanistic and Technical-Professional. In the first two years of high school the curriculum for all students is common, whereas in the last two years some differences exist.[2]

The teams in charge of constructing the tests start from a detailed curricular analysis in order to determine which CMOs and OFs are, and which are not, susceptible to assessment through pencil-and-paper tests. In addition, they must consider the orientations of organizations such as the CRUCH for the purpose of deciding aspects to include or exclude from the tests.

All of the PSU tests are formed by multiple-choice questions with five response options in which only one answer is the key. The tests admit inclusion of texts, images and other graphic resources.

The Language PSU includes the following item classes:

[2] Even though in 2009 there was a change in the national curriculum, it has not been totally implemented for any cohort presented by the examination. The change of the PSU will happen in 2014 after the curriculum has been implemented for four years. For this reason, except for some minor adjustments, currently the tests are constructed based upon the design derived from the curricular analysis done over the 1998/2008 curriculum.

- Knowledge and basic Spanish Language and Communication skills items (Section I)
- Connector items (Section II)
- Writing plan items (Section II)
- Context vocabulary items (Section I)
- Reading comprehension items (Section III)

It also makes use of alternate question formats in the Mathematics, Science and Social Sciences tests, which are: direct, indirect and combined questions.

> **Direct:** They are characterized by presenting a stimulus (variable condition, since some items directly include the question or affirmation, without the need to consider a previous stimulus) and the five options.
>
> **Combined:** They present a stimulus followed by three affirmations, the truthfulness of which must be decided, and beyond which appear the five options, which are combinations of the affirmations.
>
> **Data Sufficiency:** In these questions, the solution to the problem posed is not requested; rather, a decision is requested as to whether the data provided in the problem statement plus those indicated in the affirmations (1) and (2) are sufficient to reach that solution.

The sizes of the four PSU tests are similar:

- Language test: 80 questions
- Mathematics test: 70 questions
- History and Social Sciences test: 75 questions
- Science test: 80 questions

The Science test is formed by a common module (which assesses contents of the first two years of high school education) with 54 items: 18 Biology, 18 Physics and 18 Chemistry, and by an elective module (which assesses contents of the last two years of high school) with 26 items. These 26 questions may be of any one of the three sub-areas; the student chooses the elective test preferred.

The tests are assembled in function of the coverage of the themes and skills defined in the respective specifications tables, selecting the items from the bank one by one to cover the specifications matrix in each one of the four high school levels, while simultaneously taking into account the intended distribution of the difficulty levels that each test must have, always starting from the easiest questions and ending by the most difficult ones. Additionally, the person responsible for the test reads all of the items selected in order to verify that they are not too similar to each other. The psychometric indicators that the person assembling the test takes into account are entered into individual files for each item and are obtained thanks to the pilot application of all the constructed questions; the indicators most used in test assembly are the difficulty, discrimination and reliability, which support the development of precise and pertinent tests with respect to the measurement objective sought.

DEMRE carries out a follow up of the analysis at the item level with respect to the results of the operational test. That follow up consists of comparing the results of the operational administration with those of pilot administration. In some instances,

DEMRE uses this information to remove operationally administered items from the operational scoring process.

In the assembly process of a definitive test DEMRE seeks an item difficulty level[3] between 3% and 90%, with an average difficulty between 40% and 50%. In the particular case of the Language test, DEMRE includes a lower number of questions with high difficulty level; this has been done in response to the empirical results of previous piloting, which showed that some more demanding themes included in the past have shown to be too complex for the students. This is why, with regard to Language, the questions are in a range of difficulty between 10% and 90%. On the other hand, it so happens that in the Mathematics test there are not many questions that are very easy for the population and therefore the difficulty levels fluctuate between 3% and 80%. All of the tests are assembled by ordering the items by an increasing degree of difficulty.

In the APA standards, there is no specific recommendation about the range of item difficulties to be used on a test. Standard 3.9 states that there must be documentation of the psychometric characteristics of the test and there must be a description of how selection of the items for a test is done and the criteria taken into account to do so.

The difference in difficulty between the Language and Mathematics tests is an empirical issue that has been present since ETS's review of the PSU in 2005, where that evaluation team noted that while PSU Mathematics test was too difficult for the population of applicants, the PSU Language and Communication test showed adequate difficulty for the population of applicants. That difference has been recently exacerbated, in part, by the fact that the 2011 Mathematics test included five additional items of high difficulty in order to provide for a higher ceiling on the test to distinguish among applicants at the upper tail of the score distribution. Given the persistence of this difference over time, the PSU program should address it as part of its test construction and test scoring practices. For example, test difficult targets could be better informed by the use of item response theory as the basis of the test construction process. If this were to be done, the test difficulty could be targeted to the applicants' ability levels and maintained across years.

Another issue that the PSU program should address involves the scoring of the PSU with a correction for guessing, which rewards applicants who left unanswered questions when they are not sure about their responses. However, the computation of item difficulty is done on the basis of right and wrong responses, where omitted responses due to guessing are scored wrong. This would have the effect of increasing the apparent difficulty of the test items.

Regarding item discrimination, the data show that the average degree of discrimination of the tests is above 0.450, and in some cases over 0.600, as for the Mathematics test. Table 14 contains detailed information of the average discrimination values of each test (and where available, by form) administered until now. All of the average values are high. The test with the lowest average is Language and Communication and the one with the highest is Mathematics. "Discrimination average" corresponds to the

[3] All of the Figures and Tables included in Objective 1.1.f. show the discrimination and difficulty values for the operational and pilot items.

average of the biserial correlation coefficients when averaging the discrimination indices of each item (DEMRE technical report). The discrimination index is affected by the sample of those evaluated and by the set of all of the questions. In the specific case of the of the DEMRE admissions tests, the discrimination index is also affected by the decision of students to answer or omit a question. This decision may cause a greater tendency to answer test items that are easier than others, which leads to lower discrimination values for these easier tests and greater discrimination values for the more difficult ones. However, all the values surpass the criteria established by DEMRE itself and are adequate for the context in which they are obtained.

Table 14: Average Discrimination Values for Each Assessment

Year	Average Discrimination Values									
	Language and Communication		Mathematics		History and Social Sciences		Science Common	Science Biology	Science Physics	Science Chemistry
	FORM 101	FORM 102	FORM 112	FORM 112	FORM 121	FORM 122				
2004	0.451	0.444	0.633	0.643	0.532	0.506	0.520	0.460	0.480	0.550
2005	0.469	0.483	0.620	0.621	0.502	0.512	0.520	0.490	0.540	0.580
2006	0.502		0.681		0.516		0.497	0.459	0.551	0.556
2007	0.512		0.709		0.552		0.588	0.585	0.568	0.699
2008	0.510		0.639		0.539		0.570	0.593	0.610	0.656
2009	0.503	0.504	0.644		0.548	0.549	0.541	0.519	0.649	0.530
2010	0.487	0.498	0.644		0.547	0.541	0.533	0.514	0.550	0.535
2011	0.445		0.667		0.562		0.576	0.552	0.671	0.544

With respect to the reliability index of the tests, a report produced by the DEMRE Research and Studies Unit stated that:

> All tests, both as a full group for each of the subsets determined by the variables Region, Gender, Dependency and Graduation Year, presented α reliability coefficients at or above 0.91, a value which, according to international and national standards, is considered very satisfactory, for it indicates that 91% or more of the variance of the scores is the result of individual differences in performance presented by applicants. This result ensures that ordering of applicants is done from very precise scores, which, in turn, guarantees the quality of the selection performed, which is the ultimate end of the process. (DEMRE, 2009, p. 4)

DEMRE does not calculate the *conditional* standard error of measurement (CSEM) of the PSU, neither classical nor IRT, which could provide standard errors conditional to specific cut points. Nevertheless, the classical standard error of measurement (SEM) that DEMRE does calculate is not particularly helpful to stakeholders who wish to understand PSU scores because the classical standard of measurement calculated by DEMRE is reported in terms of raw score units that may not be appropriate due to formula scoring (i.e., correction-for-guessing) on which PSU postulation scores depend.

See Table 15. The readers will be referred to the discussion for Objective 1.1.i for to find a deeper explanation and evaluation of test score accuracy and precision.

Table 15: Standard Error of Measurement for PSU Tests by Year

Standard Error of Measurement						
	PSU Test					
	Language and Communication	Mathematics	History and Social Sciences	Science Biology	Science Physics	Science Chemistry
2006	4.58	4.69	4.33	3.85	4.36	4.88
2007	4.14	4.06	4.43	3.83	4.33	4.85
2008	4.38	4.18	4.25	4.19	4.43	4.92
2009	4.29	4.22	4.34	3.63	4.17	4.46
2010	4.12	4.12	4.17	3.73	4.18	4.24
2011	3.72	4.14	4.04	3.73	4.31	4.41

Another one of the statistical indicators produced for test analysis is the DIF, which indicates if the questions operate in a differentiated way in subpopulations other than the target group. This statistic is analyzed to ensure the fairness of the instrument for the total population assessed. The DIF analyses carried out with PSU data have shown that there are important fluctuations between the values obtained from the pilot application and those obtained during the operational administration of the PSU. This may be due to the fact that the same conditions do not exist during the pilot and operational test administrations. For example, the motivation of the student is likely to be quite different on a pilot administration than it would be for an operational administration and this could have an effect on the DIF statistics. For this reason, the decision has been made to exclude from operational test assembly only those items that have severe DIF statistics (i.e., ETS DIF classification ”C”), while accepting those with a moderate or irrelevant DIF (i.e., ETS DIF classifications ”B” and ”A”). See Objective 1.1.g. Facet 1 for further discussion of the ETS DIF classifications and DEMRE’s rules for retaining items exhibiting DIF on operational test forms.

The final assembly of every test must be approved by the members of the respective DEMRE Committee before being delivered to the printer for the printing of the plate. It must be pointed out that generally two forms of the test are printed, as a mechanism of copy control. Once the test plate is printed, the professionals in charge of the tests shall review the plate and issue their approval for the massive printing of the instrument, a process which, during the printing and distribution and booklet collection phases, is under the responsibility of the DEMRE Logistics Unit Head.

These psychometric design elements are supported by the security procedures used during the distribution and collection of the test booklets. Specific control protocols track the booklet numbers and booklet packages as they are delivered to and returned from the application sites. In addition, the police force is available to ensure the safety of the test material.

Finally, the PSU design, from its first application, produced a standardized scale with an average of 500 and a 110-point standard deviation. This same scale is still used and is the basis for the individual scores that applicants receive and provide to the CRUCH universities for the selection processes. It must be pointed out that universities are free to apply different weights to the PSU results in accordance to the entrance profiles that they have established for the training programs they offer. So the universities

establishing for themselves the cutoff points that they consider pertinent towards admitting the applicants or not.

Note about the Lack of Item Anchoring

In the review of the technical document provided by DEMRE, there is an articulated plan to include past operational items (fewer than 10, which itself is insufficient for a test of this length) in new operational tests. However, this plan has never been put into practice.

Interviews with DEMRE reveal that past operational items have been placed on pilot tests. But even here, DEMRE does not analyze these past operational items in conjunction with the newly piloted items. So no use is made of the statistical performance of these items during the pilot except to fulfill the curiosity of the test developers. No common calibration of old items with the new ones occurs; hence, there is no anchoring. Anchor items would provide a vehicle to compensate for differences in test difficulty and thus contribute to test fairness.

Note about the Lack of IRT Analyses

The IRT framework has not been incorporated for test construction and equating. DEMRE does not perform IRT analyses in the following sense. Running a statistical package to generate IRT statistics does not mean that IRT analyses have been performed, because running the program is only one of many steps. IRT difficulty and discrimination criteria are reviewed during test construction, but only as an addition to the CTT criteria, which alone determine how tests are developed.

The PSU is a classically produced test from beginning to end. It is constructed using the statistics provided by CTT, and the scores are generated using CTT analyses. IRT statistics generated by running a statistical package have no bearing whatsoever on the psychometric analysis of PSU scores.

EVALUATION

With respect to sampling, PSU design team shows the proper consideration of the characteristics of the target population. There is a basic level of psychometric knowledge among the team members and, with respect to item selection, there is support for the statistical targets from CTT and international standards (levels of acceptance for statistics indicators). That is, DEMRE duly documents CTT statistics gathered used for the construction of the PSU. However, during interviews DEMRE reported that cases exist where some indicators (e.g., average difficulty level or the discrimination index) depart from the desirable criteria. DEMRE's documentation does not account for analyzing these departures.

It is evident that even though there is documentation describing the criteria considered in the test design, it would be desirable to have the support of bibliographic technical references, as well as of studies carried out with the test data, for each one of the decisions referred to such criteria. A high-stakes test such as PSU also should include measures of precision as part of the test construction criteria. Such criteria, e.g., either a classical or IRT-based CSEM, would allow test developers to focus and minimize errors on particular portions of the score scale.

The evaluation team has evaluated the documented characteristics of the past operational items used during piloting that have been labelled as "anchors." The evaluation team has found them to be sub-optimal, i.e., below international standards. These items are in no manner whatsoever anchor items. First, although the DEMRE's documentation refers to so-called "anchor sets," in practices these item sets are not used for item calibration and score equating. Even if they were used for calibration and equating, their absolute number is so low that they would not suffice to accomplish the task in a valid and reliable manner. (See *Objective 1.3. Evaluation of IRT models for item calibration, test development and equating purposes* for a longer discussion of this point.)

Table 16 shows a summary evaluation for the design of the PSU test. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 16: Summary Evaluation of PSU Design of the PSU Test

Facet 2: Design of the PSU test	
1. Describe the specifications or guidelines used in the construction of the operational PSU form.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. What consideration is provided to the intended population of test takers when designing a test? • What consideration is provided to the intended use and meaning of the test scores when designing a test?	E
b. What content and statistical criteria are used in the test item selection? • What considerations are provided to the intended population of test takers, to the score use and interpretation, and to socio-demographic variables when selecting both content and statistical criteria? • What types of committees participated in the definition of both content and statistical criteria? o What was the process followed by the committees? o What type of training was provided to the committees?	C (2.1)
c. How was the length of the test thought out? • What considerations were provided to the intended population of test takers, to the use of scores and score interpretation, and to the socio-demographic variables when selecting both content and statistical criteria? • What types of committees participated in the definition of criteria for both contents and statistics? o What process was followed by the committees? o What type of training was provided to the committees? • How was the time length for the administration of the test defined? o What considerations are provided to the intended population of test takers, to the score use and interpretation of socio-economic variables when selecting both content and statistical criteria? o What types of committees participated in the definition of criteria for both contents and statistics? • What process was followed on the part of the committees? • What type of training was provided to the committees? • What validation research has taken place in support of using the PSU as a strength test?	E
d. What are the ideal difficulties, discriminations and content representations of the PSU tests? • What was the process followed in determining the ideal objectives? o What types of committees participated in the definition of both content and statistical criteria?	E

▪ What process was followed by the committees? ▪ What type of training was provided to the committees? ▪ What validation research has taken place in support of using the PSU as a power test? o What validation research has taken place in support of employing the PSU as a power test?	
e. What types of item formats are included in the test? • What type of research has taken place with PSU item format (for example, cognitive laboratories, tree analysis, etc.)?	C (3.6)
f. What is the target and test precision reliability? • What considerations take place with respect to the intended use and meaning of the test scores when designing a test? • What consideration was provided to the type of score use (for example, decision, social benefits) and to the type of score and scale? • What considerations are provided to the intended population of test takers, score use and score interpretation, and to the socio-economic variables when selecting the target? • What process was followed for determining the target? • What type of committees participated in the definition of the target? ▪ What process was followed by the committees? ▪ What validation research has taken place in support of using the PSU as a test of strength? • What validation research has taken place in support of using that target?	C (2.1)
g. Describe the process followed for publishing the PSU test booklets (for example, the delivery of the booklet). • What type of quality verification processes take place after the publishing of the test? • Who carry out the tests and what is the process to communicate the results?	E
h. Why are there only two forms? • What psychometric considerations are constructed when developing the forms (for example, equalization, item request, anchor items)? • What has been the role of research / policies with respect to these decisions?	C (2.1)

RECOMMENDATIONS

1. We recommend a better definition of DEMRE's test construction targets, e.g., a tolerance level for conditional standard of error of measurement. The PSU program should identify the portions of the score scale where greater precision is required and construct the test accordingly.

2. We recommend documentation of the criteria for test construction. This documentation should list
    a. participants' characteristics and qualifications,
    b. clear definitions of primary and secondary uses of test scores, and
    c. analyses of the consequences on test scores when departing from test construction criteria.
3. We recommend that anchor items be used for their intended purposes, i.e., to link forms together for the goal of facilitating calibration and equating. The PSU program should also review the criteria for selecting anchor items—including the coverage level of the specifications matrix cells or, at least, the distribution of these items throughout the thematic axes of each test—to reach international standards. We recommend updating DEMRE's anchor set specifications to comply with international standards.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2006). *Estudio de la confiabilidad de pruebas de selección universitaria. Proceso de admisión 2006.* Santiago: Universidad de Chile.

DEMRE. (2007). *Estudio de la confiabilidad de pruebas de selección universitaria. Proceso de admisión 2007.* Santiago: Universidad de Chile.

DEMRE. (2008). *Estudio de la confiabilidad de pruebas de selección universitaria. Proceso de admisión 2008.* Santiago: Universidad de Chile.

DEMRE. (2009). *Estudio de la confiabilidad de pruebas de selección universitaria. Proceso de admisión 2009.* Santiago: Universidad de Chile.

DEMRE. (2010a). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales 2011*. (Admisión 2012). Santiago: Universidad de Chile.

DEMRE. (2010b). *Estudio de la confiabilidad de pruebas de selección universitaria. Proceso de admisión 2010.* Santiago: Universidad de Chile.

DEMRE. (2010c). *Prueba de selección universitaria (PSU): Antecedentes y especificaciones técnicas.* Santiago: Universidad de Chile.

DEMRE. (2011a). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2011b). *Estudio de la confiabilidad de pruebas de selección universitaria. Proceso de admisión 2011.* Santiago: Universidad de Chile.

DEMRE. (2012). *Actualización marco teórico ciencias naturales (física).* Santiago: Universidad de Chile.

Objective 1.1.c. Facet 3. Specifications of the test construction

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for developing test specifications. A framework for evaluating PSU approaches for developing test specifications is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.3

> The test specifications should be documented, along with their rationale and the process by which they are developed. The test specifications should define the content of the test, the proposed number of items, the items formats, the desired psychometric properties of the items and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information. (p. 43)

Standard 3.5

> When appropriate, relevant experts external to the testing program should review the test specifications. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented. (p. 43)

GENERAL DESCRIPTION

Even though there is no manual for test construction, the process followed by each Committee is similar, which is evident from the theoretical frameworks employed.

In every case, test construction began with a basic consideration of the test policy proposals, issued by the CRUCH, Consejo de Rectores, which focus on the test's purpose (it is to be a selection test) and its nature (it should combine knowledge and skills, unlike its predecessor, which was an aptitude test).

Another important point of reference was the set of decisions made by the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior*, an organization established when the first operational application of the PSU was about to take place in order to establish guidelines for test construction. A key manner that the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior* has oriented test construction is by Curricular Reference, which involves linking reasoning skills and curriculum contents under fair conditions, that is, without giving priority to either component. As well, the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior* also prescribed the gradual inclusion of certain curricular contents in the test, year after year, until a test version was produced, in 2006, which covered the whole curriculum. This is an important variable which has been considered during the construction of the tests.

Additionally, it has been necessary to analyze extracurricular aspects, related to the demands made by the universities in the different fields on those entering the first semester of a career, with the purpose of making decisions regarding the inclusion of some content areas that are not expressly contemplated in the curriculum or concerning the adjustment of the content assessment strategy that is already part of the curriculum.

With this complex of variables to consider, each DEMRE Committee has carried out, regarding the construction of its own test, a curricular analysis of the Minimum Required Contents, CMO, and of the Fundamental Objectives, OF, intended for the national curriculum.

The CMOs correspond to the set of conceptual knowledge and practical performance capacities (procedure knowledge and practice), which the students are required to learn and that are defined in each curricular sector and subsector as necessary in order to reach the fundamental objectives. The contents point towards three broad categories of learning: knowledge, skills and attitudes. The Fundamental Objectives, on their part, are the competencies or capacities that must be achieved by students to complete the successive levels of High School education and that orient the whole teaching-learning process.

Also, the curriculum includes Learning Expectations (*Aprendizajes Esperados*), understood as "the north which guides teaching" in the official curriculum. These form a minimum benchmark of what the students must learn in each unit. Without limiting the possibility of expanding or deepening the aspects that they point to, these benchmarks guide the decisions regarding what weight to provide each theme (*eje temático*) or skill within the specifications matrix of a test.

In each Committee, the task of analyzing the curriculum of the respective area has implied a series of reorganizations and rankings of the CMOs and OFs, taking into account the weights or importance that these have within the curriculum, the degree of precision or generality with which each one is described and the measure in which they may be directly associated with each other. The relations between declared skills as part of Learning Expectations and the skills that form part of Bloom's taxonomy were also analyzed, selected by the UCP in order to orient the PSU construction in its cognitive dimension and to construct a general hierarchy of what should be assessed.

The Bloom's taxonomy referred to above is the Bloom, et al. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Dimension.* This taxonomy classifies cognitive abilities using the following levels:

- Knowledge
- Comprehension
- Application
- Analysis
- Synthesis
- Evaluation

A short-coming of Bloom's taxonomy is that it is "one-dimensional." That is, conceptual kinds of understanding as indicated by the Knowledge level should be distinct from cognitive processes from Comprehension through Evaluation. This "one-dimensional"

taxonomy has been revised in Anderson, et al. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* The revised taxonomy improves on the original taxonomy by providing a two-dimensional framework for classifying assessment items by Knowledge Dimension and Cognitive Process Dimension. In the revised taxonomy, the Knowledge Dimension now consists of:

- Factual Knowledge,
- Conceptual Knowledge,
- Procedural Knowledge and
- Metacognitive Knowledge.

This dimension is crossed with the Cognitive Process Dimension of:

- Remember,
- Understand,
- Evaluate,
- Apply,
- Analyze,
- Evaluate and
- Create.

Moving to the revised taxonomy would allow DEMRE to classify a broader range of PSU assessment targets than it currently does.

Each Committee also discussed which elements should be assessed could be assessed by pencil-and-paper tests, and based upon that discussion, the exclusion of some curricular aspects was determined. According to the technical team report during the interview sessions, approximately 80% of the national curriculum contents are measurable by pencil-and-paper tests (MINEDUC, 2011).

In a parallel manner, test construction attempts to take into account the fact that the national curriculum has two curricular branches: the Scientific-Humanistic and the Technical-Professional. Even though both curricular branches have a considerable amount of content in common, they differ especially during the third and fourth years of high school. (These differences pose a dilemma for the PSU, which is based on curriculum. While the national curriculum was designed to address the needs of two different groups of students, the PSU assessment frameworks were designed to address the needs of one of these two groups. Thus, the specifications matrices for the PSU came to target the third and fourth years of the Scientific-Humanistic curricular branch, while neglecting those of the Technical-Professional curricular branch.)

As the final result of the analysis carried out in each Committee, the specifications tables or curricular branches which guide question construction are produced. The curricular matrix of each test is a double entry table, out of which arise spaces or boxes in which the number of items to be constructed are defined, reflecting the ranking done with the curricular contents. Then, the tables or matrices are formed, on the one part, by the specific Thematic Axes derived from the CMOs and, on the other hand, by the Cognitive Reasoning Skills that arise from the pedagogical actions derived from the curriculum OFs. The latter is in regards to each of the four years of high school education that are intended to be covered by this assessment.

The specifications tables are one of the basic tools in test construction used by the different Committees, with which it is assured that, throughout the years, the conformation of each test is comparable in essence, with the exception indicated previously with respect to the prescription of the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior* on the gradual inclusion of certain content between the years of 2004 and 2006. These tables constitute basic working materials for the item writers who, as part of their training, come to understand their significance as construction guides.

The specification tables are published in the PSU web site; therefore, they are public knowledge.

Since the theoretical frameworks of the tests are part of the material delivered to the item construction commissions, the constructors receive in writing question samples which serve as models for the construction process. Likewise, according to the technical documentation, part of the training process includes the review of items which have been used in previous test applications and which have undergone failures, which are used to inform the constructors with regards to the mistakes common in construction and which they should avoid. Finally, the technical documentation reports that DEMRE professionals participating in the item construction commission are responsible for orienting the review workshops, where the indications are set on item elaboration guidelines. In this order of ideas, even though there is no written manual properly such for item construction, it is clear that the instructions to build them are the dominion of the technical teams and become known to the constructors through training and feedback during the construction and review workshops.

EVALUATION

The test construction process rests upon the training of the participants, the DEMRE team professionals as well as the rest of the members of the commission. Even if it is true that the good training level of DEMRE professionals supports the assurance of the process quality, it is clear that a test construction manual would be useful, as a technical and instruction-orienting document for those participating in test construction. In that way the standardization in the communication of item acceptance and rejection criteria and of the guidelines and their construction recommendations is assured. Additionally, it is worthwhile mentioning that a training process including more training time for future developers, such as the one, which according to the documentation, is carried out with respect to the Language area developers, ensures the adequate appropriation of the theoretic framework, of the test specifications and of the item construction guidelines among the constructors, and expands the opportunities to see and to analyze item examples and models of the different formats used, which is why it would be desirable to apply this same training process in the commissions of the rest of the tests.

Table 17 shows a summary evaluation for PSU test construction specifications. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled “Rating” in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 17: Summary Evaluation of PSU Test Construction Specifications

Facet 3: PSU Test construction specifications	
1. Describe the PSU test construction manual.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Provide a PSU test construction manual summary. • Point out some of the main sections on specifications.	C (3.3)
b. Who participates as author of the test construction specifications?	C (3.3)
c. Which is the review process for test construction specifications? • What criteria does the review involve? What is the closing process for blocked specifications for test construction?	E
d. How well detailed are the test construction specifications? • Who constitutes the main audience for test construction specifications? • Who constitutes the secondary audience for test	E

construction specifications? • Is the document publicly available?	
e. Are examples provided regarding the critical tasks in test construction specifications?	D*
f. Are the test construction specifications readily available for all participants during the test construction process? • What amount of training is provided to the test construction participants?	E

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. We recommend providing a manual for test construction.
2. We recommend that the test development training process be unified by generating standardized guidelines that are taught to all of them. These examples ought to illustrate mistakes that should be avoided and aspects that should be considered for achieving compliance with the established acceptance criteria. The training process should also provide enough time to verify that the new developers comprehend the frameworks and the test specifications before they start the development task properly as such.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A. Mayer, R. E., Pintrich, P. R., et al. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Completed ed.). New York: Longman.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The cognitive domain.* New York: Longman.

DEMRE. (2010a). *Descripción técnica de la prueba de matemática.* Santiago: Universidad de Chile.

DEMRE. (2010b). *Marco teórico prueba de selección universitaria historia y ciencias sociales.* Santiago: Universidad de Chile.

DEMRE. (2010c). *Prueba de selección universitaria (PSU): Antecedentes y especificaciones técnicas.* Santiago: Universidad de Chile.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2012a). *Marco teórico de la prueba de selección universitaria (PSU) del sector ciencias naturales subsector de biología*. Santiago: Universidad de Chile.

DEMRE. (2012b). *Marco teórico prueba de lenguaje y comunicación.* Santiago: Universidad de Chile.

DEMRE. (2012c). *Marco teórico PSU-ciencias-química.* Santiago: Universidad de Chile.

DEMRE. (2012d). *Actualización marco teórico ciencias naturales (física).* Santiago: Universidad de Chile.

MINEDUC. (2011). *Aprueba bases administrativas, bases técnicas y anexos de licitación pública, sobre servicio de evaluación de la Prueba de Selección Universitaria (PSU)* (ID Nº 592-44-LP11). Santiago, Chile: Autor.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.c. Facet 4. Specifications matrix

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU specifications matrix. A framework for evaluating the PSU approaches for reviewing the specifications matrix is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.3

> The test specifications should be documented, along with their rationale and the process by which they are developed. The test specifications should define the content of the test, the proposed number of items, the items formats, the desired psychometric properties of the items and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information. (p. 43)

Standard 3.5

> When appropriate, relevant experts external to the testing program should review the test specifications. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented. (p. 43)

GENERAL DESCRIPTION

Each PSU test has a matrix or specifications table that was formed as the result of an analysis carried out in each Committee on the curricular provisions present in the four grades of high school: specifically, on the Minimum Required Contents (CMOs) and the Fundamental Objectives (OF) proposed in the curriculum. (Note that this refers to the declared National Curriculum rather than curriculum as implemented in secondary classrooms.) These analyses were the responsibility of the teams of professionals of the DEMRE committees, all of whom have university training in the field and postgraduate degrees in the subject area and/or in measurement and assessment. With that extensive training, the team has the capacity to perform curricular analysis of each area is ensured, the ability to decide what curricular aspects that may be assessed through a pencil and paper, widely distributed application test, whose purpose is to select students for entry into higher education. Though that is the case, there is no evidence that the decisions taken by these teams are validated by outside reviews of people expert in media curricula, as well as in the training demands imposed by the universities which use the test as selection criterion.

For all tests, the specifications matrix is a double entry table whose columns and lines define themes (*ejes temáticos*) and cognitive skills (*habilidades cognitivas*) that are the object of the assessment; this matrix also presents the number of items in each dimension. These item numbers reflect a previous ranking of the curricular contents of high school (especially if these form an explicit part of the Expected Learning proposed

from the curriculum). They are also relevant with respect to entry into the university, aside from the practical consideration if they are assessable or not by means of pencil-and-paper tests.

The procedure for the production of each matrix is described in the theoretical framework of the respective test.

Language and Communication
In Spanish Language and Communication, the curricular analysis for the specifications matrix considered several categories (DEMRE, 2012c, p. 20):

- Sequence: oriented to focus upon the development of contents and associated skills throughout the time indicated in the curricula
- OF / CMO relation: estimation on the achievement pertinence of the OFs, in relation to the CMOs
- Skills: explicit relation of the skills in relation to the OFs and CMOs

Mathematics
The specifications matrix of the Mathematics test considers the skills which in the curriculum are posed as mainly associated to the learning of Mathematics (DEMRE, 2010a, p.4).

• **Procedures which may be standardized:** includes the development of skills that are put in place for the learning of different procedures and methods that enable the fluid use of instruments, the carrying out of calculations and estimations, the application of formulas and conventions, which later become part of routine and algorithmic procedures.

• **Problem solving:** includes the development of skills such as identification of unknowns and estimation of the order of their magnitude, search for and comparison of solution routes, data and solution analysis, result anticipation and estimation, trial and error systematization, model application and adjustment and conjecture formulation.

• **Mathematical concept structuring and generalization:** includes the development of skills such as particularization, generalization, search for patterns and regularities, knowledge integration and synthesis, logical argument linking, and distinction between assumptions and conclusions. Also, the relations between different issues and concepts are incorporated as well as some background information relative to the historical evolution of some of them.

In parallel, reviews were completed of the OFs provided for each one of the teaching grades of high school education and the CMOs that were grouped into theme axes.

Science
Regarding the Science test, the specification table production procedure involved the review of the relations between the OFs and CMOs of the curriculum and between these and Bloom's Taxonomy, which was assumed as a reference point for assessing the cognitive dimension of the test. So, the matrix design phases can be described in the way that they occur for Biology, which, in general terms, can be assumed as

representative of the matrix construction procedure in Science. These phases include the analysis of (DEMRE, 2012b, p. 18):

- **Relation between OF and CMO**

  The theoretical framework of the subsector was analyzed to establish the relation between the OFs and CMOs. For that purpose, the information was sorted in tables by level, in which the CMOs were identified (theme axes), establishing which OFs were directly associated with a specific content and which were more general in character.

- **OF/CMO/PSU skills relation**

  The relation was established between the OF with a specific character, the theme axes (CMOs) and the cognitive skills measured in the PSU. For that purpose, the declared skills were classified in the Expected Learnings of the Plans and Programs according to the theme axis in the PSU skills taken from Bloom's Taxonomy. In the case of the first year of high school, these skills correspond to those declared by subunit for each theme axis. This was carried out for each high school grade and for each theme axis. In the case of the OFs that are general in character and not directly associated to a CMO, the cognitive skills were detached from them, establishing them as a measurable dimension to be assessed by a PSU instrument.

Starting from the aforementioned, the theoretical percentage of each one of the skills and contents to be assessed by the PSU Science test was established, from which a representative specifications table of the curriculum of each subsector forming part of the test (Biology, Chemistry and Physics) was elaborated.

Additionally, extra-curricular aspects were considered, such as content relevance in the undergraduate university environment, the population configuration on which the test is applied and its response to pilot tests.

Regarding the Social Sciences test, the corresponding analysis of the relations between OFs and CMOs were also carried out, establishing that it is possible to infer three great dimensions proper to the study of History and Social Sciences, based upon which performance indicators are written for each one of the theme axes (DEMRE, 2010b, p. 22):

- **Comprehension of social reality**: appeals to the development of a "comprehensive" view of social reality and the surroundings; the comprehension of different historical, social, political, economic, geographic and cultural processes, as well as the development of skills such as reflection and analysis of History and Social Sciences problems.

- **Civic-democratic participation**: aims at the knowledge and comprehension of the rights and duties "implied" by life under democracy, analysis and discussion concerning pluralism, human rights, civic-citizen participation, among multiple aspects. As well, includes the notion of commitment for participating in a democratic society, along with becoming involved in solving the problems and in the defense and respect of the essential rights of every person.

- **Valuing, respect and sense of belonging**: involves the knowledge and comprehension of the different social realities, of the different cultures, of different human groups, as well as the development of an "attitude" of respect in front of "historic-cultural" diversity, tolerance facing different points of view, respect for and defense of the environment, as well as the commitment and sense of belonging to a particular cultural reality, of solidarity with different communities and the protection and care of the echo system.

These dimensions are matched in the specifications matrices with three cognitive skills that are associated to Bloom's Taxonomy categories (DEMRE, 2010b, p. 27): (1) recognition skills, (2) comprehension and application skills and (3) analysis, synthesis and assessment skills.

It must be pointed out that for all areas, the specifications tables are one of the basic tools for the construction and assembly of the respective tests. These tables constitute the basic working material for the item writers who, as part of their training, are informed about how to apply these tables as construction guides.

Furthermore, the specifications tables are published on the PSU site to make them accessible to the general public.

EVALUATION

According to what was reported during the technical team interviews, the test places greater emphasis upon the Scientific-Humanistic curricular branch of Chile's national high school curriculum than upon the Technical-Professional curricular branch. It should be pointed out that the level of alignment of the matrices with respect to that the implemented curriculum, taking place in the classroom, is not known. But, once again, it would have to be verified.

With respect to the development of the specifications matrix, as per Standard 3.3, DEMRE has documented test specifications, or theoretical frameworks, for each subject area test. However, those test specifications have not been reviewed, as per Standard 3.5, by "relevant experts external to the testing program." So, although the matrices may well stand up to external scrutiny, that possibility needs to be verified. See 2.b in the table below.

Table 18 shows a summary evaluation for the PSU specifications matrix. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 18: Summary Evaluation of the PSU Specifications Matrix

Facet 4: PSU specifications matrix	
1. Describe the specifications matrix guiding the PSU test construction.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Which are the criteria with respect to the construction of the specifications matrix? • How was the intended use of the test scores incorporated (that is, selection and scholarships)? How was the intended meaning of the test scores incorporated (CRT and NRT)?	E
b. How was the specifications matrix developed? Please point out the main process components and the main actors.	C (3.5)
c. What are the main qualifications of the matrix developers?	E
d. Describe the alignment of the matrix with teaching and learning in the classroom. • Has the research been put in place with respect to the inspection of the classroom alignment throughout the regions, SES, curricular branch and type of school (municipal, private and subsidized)?	C (3.5)
e. Is it easily accessible for the participants involved in the construction of the PSU tests? • Is it operationalized?	E
f. Describe the design details of the PSU specifications matrix. • Has it changed over time? Please, point out the main changes. Describe the root causes for the changes. Describe the closing process with respect to the adoption of changes and the main actors.	E

g. Describe the alignment of the PSU design with the national curriculum. • Describe the main components of the process and the main actors. Describe the closing process.	E

RECOMMENDATIONS

1. Concerning the implementation of new specifications tables, given the 2009 curricular change, we recommend introducing a validation process of the respective specifications tables with teams of high school education and higher education experts (first semester) to add external validity to the process, emphasizing aspects such as pertinence and relevance of the aspects included in such tables.
2. We recommend subjecting the decision to place more emphasis on the Scientific-Humanistic curricular branch than on the Technical-Professional one to outside validation and to include in the theoretic frameworks of the tests the justification for this decision.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2010a). *Descripción técnica de la prueba de matemática.* Santiago: Universidad de Chile.

DEMRE. (2010b). *Marco teórico prueba de selección universitaria historia y ciencias sociales.* Santiago: Universidad de Chile.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2012a). *Actualización marco teórico ciencias naturales (física).* Santiago: Universidad de Chile.

DEMRE. (2012b). *Marco teórico de la prueba de selección universitaria (PSU) del sector ciencias naturales subsector de biología*. Santiago: Universidad de Chile.

DEMRE. (2012c). *Marco teórico prueba de lenguaje y comunicación.* Santiago: Universidad de Chile.

DEMRE. (2012d). *Marco teórico PSU-ciencias-química.* Santiago: Universidad de Chile.

Objective 1.1.c. Facet 5. Process for the construction of the PSU operational test form

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for constructing the PSU. A framework for evaluating the PSU approaches for constructing the PSU is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.6

> The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses, and when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences and demographic characteristics of expert judges should also be documented. (p. 44)

GENERAL DESCRIPTION

In accordance with the technical documentation of the Test Construction Unit (UCP) of DEMRE, operational test construction involves the use of piloted items, loaded into the Item Bank, for which there is associated psychometric information.

The assembly of each one of the tests is the responsibility of the leadership of the respective committee. The leadership of each committee forms the professional teams that take charge of the process. These DEMRE Committees consist of between three and five persons, all with professional training and specializations, as follows:

- Language and Communication: linguistics, literature, curriculum and assessment
- Mathematics: numbers, proportionality, algebra, functions, geometry, probabilities and statistics, curriculum and assessment
- History and Social Sciences: History of Chile, the Americas and Universal, curriculum and assessment
- Science: Biology, Physics and Chemistry, with the different specialties proper to each theme axis by discipline

The test assembly takes place normally between the months of April to June each school year, to be applied at the end of November or at the beginning of December of the same year. This assembly assumes that the previous stages of preparing the Item Bank have been met, such as (DEMRE, 2011a, page 22):

- Qualitative and quantitative analysis of the results from the previous year pilot application

- Tracking of types of items, according to the test section (Language)
- Tracking of types of texts and associated questions (Language)
- Item classification according to theme axis
- Item classification according to cognitive skill
- Item classification according to statistical difficulty
- Item classification according to discrimination
- DIF detection, that is, to check if the item shows bias

Generally, the assembly process takes place in the following manner:

After examining the contents of the Item Bank, the committee leader, along with one or two of its members, together make an initial selection of items that fit the test matrix requirements. These items are printed and sorted by difficulty to facilitate their organization into each of the test sections. Each file has the statistical and descriptive data of the item, which facilitates this pre-assembly work.

After which, another member of the Committee on Qualitative Aspects uses the curricular reference to evaluate the relevance of the selected items for the different test sections, the question formats, etc. If edits are suggested, the committee member who initially pre-assembled the test makes the appropriate adjustments, assembles the digital format directly into the Item Bank and prints the complete test. The complete test is then reviewed by the President and Technical Advisor of the item construction commission, who are expert academicians in the field, coming from the Universidad de Chile faculties, with whom DEMRE maintains certification agreements. Both reviewers certify the curricular reference of the questions, their correspondence with the theme axis and their relevance to the purposes of the instrument, leaving their comments in writing on a form designed for this purpose.

In case one or more items are rejected, the test form is returned to the committee so that replacements of the required items will be made and, once again, the relevant reviews are performed. The additional review is performed by another member of the committee of the respective area. At this point, the review is dedicated to the detection of possible problems in formatting (e.g., typographic, spelling, format, printing, etc.).

Having completed all of the adjustments requested in the latest round of reviews, the committee head (or his/her subordinate) takes charge of the final test assembly in the item bank, thereby creating the final edition that is sent for printing.

EVALUATION

According to the reviewed documentation, the test assembly process conforms to test specifications respecting item quantities for each area of the specifications matrix and incorporating items that comply with the statistical criteria established as acceptable. Since the test assemblers are members of the committee who have participated in the design of the test, their informed criterion is ensured in performing the selection of questions that respond to that intended to be assessed. Additionally, the review of an expert from the university environment as final reviewer of the assembled test adds safety regarding the pertinence of the instrument in a university selection process. The review process could be enriched if it included a final reviewer representing high school

education who knows the target population closely (a teacher of this educational level that has not participated in the question construction process to ensure its independence and objectivity) to validate aspects such as question clarity for the students and question pertinence in front of the curriculum carried out in the classroom.

Concerning the assembly system, even though it is partially automated, with a system that pre-selects items in function of given criteria, according to the description found in the technical documentation, a large part of the selection process for assembly is largely manual, which reduces the efficiency of the process (requires more time and human resources to have an assembled test). A more automated assembly system would reduce the risk of item duplication in one same test and the interference of subjective criteria when a choice has to be made of one among many items with similar possibilities of completing an assembly. Such a system would be greatly benefited by the involvement of the IRT framework in lieu of the CTT framework, currently used for the PSU. An IRT framework would allow targeting the tests to applicants' levels of ability in a systematic way. From a previous evaluation of the PSU, ETS reported a disproportionate difference in difficulty of the PSU test due to the lack of a test construction target for applicants' level of ability (Educational Testing Service, 2005).

Table 19 shows a summary evaluation for the PSU process for the construction of the PSU operational form during test construction. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment

standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 19: Summary Evaluation of PSU Process for the Construction of the PSU Operational Form during Test Construction

Facet 5: Process for the construction of the PSU operational form during test construction	
1. Describe the process and criteria followed with respect to the construction of the PSU operational form.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Describe how the test design is posted during the test construction process. • Describe the content and statistical criteria informing the construction of the PSU operational form.	E
b. Which are the qualifications of the participants who construct the PSU forms?	E
c. How are the participants trained or prepared in anticipation of the test construction?	D*
d. What tools / processes are utilized? • Does it consist of an automatic process? • What algorithms are used? • Which function optimizes? How are the items based upon passages allotted? How is the intended use of test scores posted during the construction of an operational form? Why has there not been any consideration of using the CSEM to inform test construction? • How is the intended population of test takers posted during the construction of an operational form? Why has there not been use of multiple test objectives with respect to the intended populations?	C (3.6)
e. Which is the time frame for the construction of an operational form of the PSU?	E

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. The test construction process has been well described within the CTT framework. We recommend that the training be formalized and documented, viz., training in test construction tools and processes to develop statistical targets.
2. We recommend considering automating test assembly to avoid the security risks that might arise in the future from the continued physical handling of the

booklets, e.g., those that arise from the repetition of questions or from inconsistencies found within the specifications table.

3. It is suggested to include a reviewer of the assembled test coming from the high school education level, contrasting with the reviewer coming from the university level.
4. We recommend a transition into the IRT framework for test construction. This transition would better position test construction activities to target the PSU tests to the applicants' level of ability in a systematic way. The IRT framework would also provide more precision, and, hence, reliability at points on the PSU scale where the important decisions are made.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2011a). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales 2011*. (Admisión 2012). Santiago: Universidad de Chile.

DEMRE. (2011b). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

Educational Testing Service. (2005). *Evaluación externa de las pruebas de selección universitaria (PSU)*. Princeton, NJ: ETS Global Institute.

Objective 1.1.c. Facet 6. Process and criteria regarding the review and approval of a constructed PSU form

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process to review and approve the PSU forms. A framework for evaluating the PSU approaches for reviewing and approving the PSU forms is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.6

> The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses, and when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences and demographic characteristics of expert judges should also be documented. (p. 44)

GENERAL DESCRIPTION

The four stages for the review of an operational test are the responsibility of professional personnel: the committee head of the area under evaluation, different members of this committee and faculty from the Universidad de Chile. (Note: DEMRE is a unit within the Universidad de Chile.) In addition to their professional and postgraduate training in their respective fields, each committee member has some degree of training in measurement and assessment processes.

The review stages begin with the pre-assembly of the test, carried out by one of the members of the respective committee. In this review stage, the following aspects are considered:

- Compliance with the specifications in accordance with the specifications matrix, in its different sections, axes and skills
- Compliance of the statistical indicators within the acceptance ranges established by DEMRE
- Curricular and discipline relevance of the items
- Editorial adequacy

The review of the compliance with the specifications is a core responsibility of each test committee. The reviewers draw on the specifications table, referencing the number of items to include per each of the cells. The review by peers of the committee prevents the inclusion of items that do not correspond with the established psychometric criteria. Once an initial test form is constructed, the reviewers examine the formal aspects of that test before approving its final version.

The psychometric criteria for the operational test include the following elements:

- Difficulty (*p*-values): ranges from 3% to 90% (or from 10% to 90% for Language and from 10% to 80% for Mathematics)
- Discrimination (biserial): minimum acceptable is 0.250
- Empty distractors: with a minimum of 1%
- DIF:
    - Mantel-Haenszel Chi-Square (MH CHI): with a maximum value of 5.02
    - Mantel-Haenszel Common Log-Odds Ratio (MH LOR): If the value is positive it favors the reference group; if it is negative, it favors the focus group.
    - Standardized Mantel-Haenszel Log-Odds Ratio (LOR Z): The item must be in the range of -2.0 – +2.0
    - Breslow-Day Chi-Square (BD): Maximum range 5.02 (generally it is coincident with Mantel-Haenszel Chi-Square)
    - DIF: ETS classification *irrelevant (A)* or, in a few cases ETS classification *moderate (B)* (items with an ETS *severe* DIF (C) are not included).
- IRT:
    - Item discrimination (a), where $a \geq 0.6$ (acceptable item);
    - Item difficulty (b), where $-5 \leq b \leq 5$ (DEMRE, 2011b)

All of the items conforming to an operational test shall comply with these requirements, which is verified by the professional from the committee responsible for the pre-assembly, as well as by the head and the other member of the committee who performs the pre-assembly review.

For their part, faculty from the Universidad de Chile have as their main mission determining the relevance of item content in terms of what demands universities make on students entering the first semester. There is no evidence in the reviewed documentation that the collaboration of outside reviewers from different universities or from different geographic regions is used.

EVALUATION

The assembled test review and approval process includes the participation of different persons at different moments, which ensures the criteria contrast which is important for endowing the process with objectivity. Personnel participating in these reviews have knowledge in the areas assessed as well as with regards to psychometric fundaments, knowing the theoretic frameworks and test specifications, thus ensuring correspondence between the test assembled and the design of the same. However, the review external to the committee members corresponds to a Universidad de Chile academician, whose participation is undoubtedly important in rendering an account of the pertinence of the test in so far as its purpose for university selection, but which could be very well complemented with the participation of a reviewer representative of

high school education, who could contribute judgments on test adequateness on issues such as the appropriateness of the language used in the questions for the students and that which is pertinent about the issues approached in the same, with the educational reality of high school education classrooms.

The psychometric criteria for the item difficulty established in general correspond to internationally accepted criteria -3 to +3. However, the evidence we have collected in our demonstration of the IRT processes for the PSU test (see Figure 18: Test Characteristic Curve for 78-Item Language and Communication Test) showed a range of ability closer to the range -5 and +5. In order to judge which of these two ranges is more appropriate for test assembly, it would be necessary to document how many questions exist with difficulty levels that are closer to and further away from the extremes of this range. The same observation applies for the rest of the defined criteria.

The psychometric criteria for item discrimination used by DEMRE on the PSU calls for an IRT discrimination parameter (*a*) to be at least 0.6. In de Ayala's text (2009, p. 101), he states that "reasonably 'good' values of [*a*] range from approximately 0.8 to 2.5." The experience of the evaluation team is that a low value of either 0.6 or 0.8 for item discrimination is defendable in practice. It is the number of low discriminating items that needs to be guarded against, not the existence of any one item in the range 0.6 to 0.8 that is problematic during the construction of a test. Ultimately, it is the distribution of the discrimination values in concert with the difficulty values that is most important for the proper construction of a test. If it were the case that DEMRE had chosen a predominant number of items with relatively low discrimination values, then that would not be in line with best practices. However, the evaluation team examined the discrimination values of the items during piloting and operational administration and found that they vast majority of them had discriminations far above 0.6. (See Figure 4 in Objective 1.1.f.)

Table 20 shows a summary evaluation for process and criteria regarding the review and approval of a constructed PSU form. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 20: Summary Evaluation of Process and Criteria regarding the Review and Approval of a Constructed PSU Form

Facet 6: Process and criteria regarding the review and approval of a constructed PSU form	
1. Describe the processes with respect to the review and approval of a constructed PSU form.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. What types of review / approval belong to the PSU test process of construction?	E
b. How many evidence reviews take place?	E
c. What are the qualifications of the evidence constructing personnel?	E
d. What psychometric criteria are used for test review / approval? • How is the intended use of the test scores during the construction of an operational form posted? I hear no talk about the use of CSEM informing test construction. Why? • How is the intended population of test takers during the pulling of an operational form posted? I hear nothing about the use of multiple test objectives with respect to the intended populations. Why?	E
3. Describe the feedback used in the test construction process.	Rating
a. What review process is used with respect to considering the review recommendations?	E
b. What is the demographic composition of the reviewers or committees? What type of information is submitted to them?	D*
c. What is the professional composition of the reviewers or committees?	E

***A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. We recommend documenting in greater detail the treatment of the drafts or testing material changed during the successive review processes.
2. We recommend that the outside reviews for the operational test review process represent a larger institutional diversity (that is, for not all of them to be exclusively from the Universidad de Chile).
3. We recommend documenting more extensively the procedure to follow when one of the outside reviewers suggests eliminating or replacing an item from a pre-assembled test.
4. The recommendation is to document in a precise way the instructions on test assembly, indicating the ideal distribution of questions in function of the statistical indicators that are taken into account, that is, the maximum and minimum acceptable item number with a certain discrimination level, etc., in order to ensure test comparability between different applications.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Gilford Press.

DEMRE. (2011a). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales 2011*. (Admisión 2012). Santiago: Universidad de Chile.

DEMRE. (2011b). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

Objective 1.1.d. Quality standards in item bank management

Effective maintenance of the item bank should address issues of data management, version control, adequate identification of question availability, and replenishment strategies.

The evaluation team developed and performed interviews with relevant stakeholders from DEMRE on March 22, 2012. The interview process took a total of two hours from 9:30 to 11:30. The purpose of the interview was to gain deeper understanding on:

- Item bank structure (e.g., logical design, platforms, fields, and records) (Facet 1)
- Item bank tools (Facet 2)
- Security access protocols and processes (Facet 3)
- Process flow for updating and adding records to the item bank (Facet 4)

All the interviews were performed within DEMRE offices following an agreed-upon schedule for the visit. The interview covered the four facets and relevant elements as agreed upon with the TC during the goal clarification meeting in Santiago, Chile, in January 2012.

The following DEMRE staff participated in the interviews:

- Head of the department of test construction
- Coordinator of test construction committees
    - Mathematics
    - Language
    - History and Social Studies
    - Science

- Head of research unit and his team
- Head of information
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees.

The following subsections contain the results of the evaluation for Objective 1.1.d., Facets 1-4.

Objective 1.1.d. Facet 1. Item bank - Structure

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU item bank structure. A framework for evaluating the PSU approaches for the item bank structure is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

There is no specific standard for this objective; reference is made to general recommendations of the standards targeted to safeguard the confidentiality of the information with subjects related to the quality of the assessments such as validity and reliability.

GENERAL DESCRIPTION

The item or question bank has two main components: the item bank itself and the associated Safe Question software. The second component allows the question of elaboration or construction "from outside the bank," that is, it is a software which may be installed on any computer by the item writers.

The banking system includes:

- A server that hosts the database engine (and the database itself)
- A server that has the programs (applications)
- Mainframe computers having the security systems proper to them
    - User computers
    - An associated program: "Safe Question" (this is used in item construction)
- A single database including all of the item and test information (assemblies)
- The Safe Question software, which requires Microsoft Office Word.

The Safe Question software enables the elaboration of the items, including all of their components: statements, answer options (distractors), related graphics and text (as in the case of a Language test). The elaborated items are saved in files predesigned for such purpose and are included in the item bank. Each one of the users has precise and clearly described functions.

The item or question bank is based upon an ORACLE architecture. This bank is formed by a series of modules or programs (applications) developed in "ORACLE 6i." Each program performs a series of tasks to a database. The database is installed on an Oracle 9i database server. The related applications allow for the revision of items as well as the assembly in the official tests that are administered to the students.

The new items, incorporated with the files generated by the Safe Question software, are considered to be raw items that will undergo various processes in order to be improved and qualified, prior to a decision to incorporate them into a pilot test or the official tests.

From this point of view, the information fields for each one of the items, which may be divided into three groups, are very important:

- Information related with the theme or educational content of the item: the field, the educational grade, the specific theme content made reference to and that

which corresponds to the official curriculum; in other words, it is a direct transcription of the curricular framework, including the cognitive field
- Information related to the item characteristics such as the correct answer, the difficulty level and related texts
- Information related to the item writer

At the moment of loading the items into the system, the item is allotted a code and the system verifies that this code has not been allotted previously in order to avoid information duplication..

After its inclusion into the Item Bank, additional information is included, such as:

- Item review process: dates and participants of the changes carried out on the original item
- Item condition

This latter information is one of the most important ones because it clearly defines what has happened and what is happening with the item. For example, it states if it has been approved or not, if it has already been piloted or pre-tested, if it has some modification, if it is ready for assembly, if it was rejected during piloting or if it has been applied once or more times, all of which allows a detailed follow-up of what has happened with the item.

After the item has been piloted, its associated statistical information is included, making reference to:

- Index of difficulty (relative frequency) (correct answer and options and variables)
- Index of discrimination (correct answer and options and variables)
- IRT parameters *a* and *b*
- Item ICC Graphics
- Information function
- DIF by gender and dependency variables

The system is able to generate item use statistics and information concerning any changes an item has undergone since its inclusion in the bank.

EVALUATION

The documents reviewed contribute general and apparently complete information on the item bank. However, there are several specific issues about which no mention is made. One of them is that in relation to the modular structure of the bank or of the software representing it. Even though that which the bank does is mentioned, there is no detailed information on its structure enabling a clear judgment on its architecture.

How the bank is organized and the interactions between those operating it and the software are clearly understood. The information present is more of a look from the perspective of systems engineering than from the perspective of psychometrics, where those relations would take on meaning in relation to the assessment process in the educational context and which happens to be important for this process. The organizational relation between the Safe Question software and the bank is not established explicitly.

In spite of having a clear structure and a powerful database with much information, there is no mention if statistical information related to the bank use is being produced. The psychometric use of this type of information could guide the PSU developments in the immediate future and in the medium term. It is worthwhile to develop reports which inform the designers on test behaviors and not only on the use of items and their statistics.

There is no mention in the documents on the criteria used in the definition of the bank update beyond the item quantity by region of the specifications table.

Table 21 shows a summary evaluation for the PSU item bank. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled code in the table, usually there would be a list of professional standards not met is shown within parentheses. However, because there is no specific standard for this objective, no such listing will be made.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 21: Summary Evaluation of PSU Item Bank

Facet 1: Item bank	
1. Describe the structure of the item bank (platform, fields and records).	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Describe the logic design (module number and their relations).	D*
b. What is the organizational structure of the item bank?	F
c. How are the related items identified? (for example, base items of passages)	E
d. What use indicators are there in the item bank? (for example, DNU standards, item exposition rates, withdrawal rates)	C**
e. What other variables are important?	F
f. Which computer platform lodges the item bank?	E
g. What criteria are used with respect to updating the item bank information?	D*

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

**Because there is no specific standard for Objective 1.1.d. Facet 1, no parenthetical listing of "Standards not met" will be made for a rating of "C." This rating was based instead on the evaluators' understanding of commonly used practices in the assessment industry.

RECOMMENDATIONS

1. Insufficient technical information was found describing the item bank beyond that related to the architectural base. We recommend supplying the information that is missing regarding its modules, their functionality and characteristics.
2. Even though the aforementioned aspects of the item bank are clearly laid out, we recommend the production of more precise technical characteristics of the test elaboration process; specifically, technical and use manuals need to be generated to facilitate an understanding of what takes place and the real scope and limitations of the bank.

3. Although there is no specific standard for what indicators should be included in an item bank, we recommend additional item use indicators be added to the item bank. For example, knowing the administration history of an item would allow us to calculate its exposure rate.

BIBLIOGRAPHY

DEMRE. (2011). *Banco de ítemes.* Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.d. Facet 2. Item bank - Tools

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU item bank tools. A framework for evaluating the PSU approaches for the item bank tools is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

There is no specific standard for this objective; reference is made about general recommendations on the standards oriented to safeguard the confidentiality of the information with issues related to the quality of the assessments, such as validity and reliability.

GENERAL DESCRIPTION

The software used for the storage and tracking of the PSU test items has been designed specifically by DEMRE in an ORACLE architecture, based upon the Client/Server model. The Information Technologies Unit of DEMRE is in charge of carrying out the technical as well as operational maintenance. The System Administrator has access to all of the information of the system, except for the item statements. The Collaborator is exclusively responsible for the Safe Question software.

There is a user manual for the Safe Question software that includes information on the technical characteristics that must be taken into account when elaborating on the questions and that must be considered during committee review. However, there is no mention of a manual for the operation, versus the use, of the Safe Question software or the Item Bank.

The Item Bank platform operates as an offline internal network in such a way that the information contained in it only may be accessed internally. The access requires a password or code allotted to each one of the authorized users, each of whom has an assigned level of access to information.

The authorized users belong to three groups or administrative units:

- Information technologies
- Test constructors
- Studies and research

As it was previously mentioned regarding the information technologies unit, there are only two roles, that of administrator and that of collaborator, with the previously mentioned accesses to the system.

Regarding the group of constructors, the roles are those of unit head and coordinator, committee head and committee member. This group of persons has direct access to the items, their conditions, and to the assembly in the official tests.

The members of the studies and research unit have access to the bank but only to that which is related to statistics.

The item modifications are directly controlled by the item bank; that is, on some occasions, the software allots the condition, and in others, the person in charge does so. When done manually, a rigorous procedure has to be followed to avoid problems.

EVALUATION

The documentation describes the bank and its use clearly and completely, from a perspective more or less on the use of the same, but documentation that is much more technical is missing, allowing the assessment of the tool from a system perspective. There is no information related to the flexibility of the bank in relation to the incorporation of changes starting from possible substantial transformations in the case of changes in the PSU assessment process, as for example, changes in the test structure, item format or models or methods used in item calibration or production of results.

Neither is there mention of the bank capacity, or if this is unlimited, it is likely that someone responsible for its handling know this, but it is possible that no consciousness exists on its possible limitations.

Table 22 shows a summary evaluation for PSU tools (software and database). The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 22: Summary Evaluation of PSU Tools (Software and Database)

Facet 2: Tools (software and database)	
1. Describe the PSU item bank database tools (software and database).	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Describe the source of the software/database. • Is this software original or purchased? • Upgrade policy? • Technical backup?	E
b. What type of documentation is there (User Manuals)?	E
c. Describe the degree of flexibility incorporated into the platform. • Could you give me an example? • How do you see the future with respect to adding changes to the system, and how much flexibility does the system have incorporated in order to fit the changes? For example, bank storage of reactive items that measure writing skills.	D*
d. How do you control or protect the item versions?	E
e. What is the storage capacity and size of the item bank?	D*
f. Is this platform a server online or local?	E
g. How do you allocate the authorized users?	E

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. We recommend a technical document describing the total technical features of the software.
2. We recommend a clearer description of the specific rules for how user-level allocations are made.

BIBLIOGRAPHY

DEMRE. (2011). *Banco de ítemes.* Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.d. Facet 3. Item bank – Access security

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for accessing the security of the item bank. A framework for evaluating PSU approaches for accessing security of the item bank is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

There is no specific standard for this objective; reference is made to general recommendations on the standards oriented to safeguard the confidentiality of the information with issues related to the quality of the assessments such as validity and reliability.

GENERAL DESCRIPTION

It is worthwhile mentioning again that the system is offline in a local network; that is, it may only operate through computers that are in the network with the item bank server and on the site specified by DEMRE.

The information technologies access is established based upon user profiles which allow the allotment of a user account with its respective password. The physical access is restricted to differentiated working areas according to the test.

In general, there are two processes that enter information from outside: the first one has to do with the Safe Question software, and the second one with the allocation of statistical values after processing the IRT and DIF data. In addition, much of the information comes from data tables (IRT and DIF). Just the same, the Safe Question data may be sent from e-mails; therefore, it would be important to know the way in which the system is updated with this information.

The access to the database is differentiated depending upon each user's profile, and a hardware key is required. In general terms, the test construction unit, the head of the unit, is responsible for process validation, allotment of privileges and data updating. The persons in the test construction unit have privileges regarding item construction, editing and maintenance; they have this access using a hardware key which has been allotted to them and which is personal. From the point of view of the physical location, each test committee has access to a specific space, and from the information technologies point of view, they do not have access to the questions of other committees.

The database including all of the item information is located in a server, which has an engine for managing it.

In relation to test assembly, the authorized user validates this assembly, allotting an intermediate condition or final confirmation of the form, in order to prevent it from being edited. For security reasons, once a final confirmation exists, any possibility of modifying the items in real time is disabled. If a modification is attempted, this is observed only until the item has been released from the final confirmation restriction. This action automatically performs the change in the conditions of the questions. A question's content, as well as its correct option, are stored in the database in an encrypted form. Only the correct options to each question are decrypted once the test has been applied. This action is carried out by the head of the Test Construction Unit.

In general terms, the system prints output to text format or exports documents to Excel, allowing for the creation and management of tables on personal access computers, as required.

Since several processes are performed manually, the verification of their correct performance must be carried out by the head. Furthermore, the system has internal validation processes in order to verify that the processes are carried out correctly.

EVALUATION

In the information provided there is some mention to aspects such as system protection, backup systems for the information and system auditing. The information on the data banks has no backup in other external files, including, for the unit handling the bank, increasing the risk of total loss of the information due to an accident. There are no procedures or resources assigned for the protection of the system from possible penetrations due to computer viruses, since no information flows through pen drives or USB.

There is no programming for updating and maintenance of the bank in the sense of including new technologies which enable its development and pertinence in relation to PSU changes (number of assessed, need to update the tests due to changes in the educational reality, etc.).

Table 23 shows a summary evaluation for PSU tools access security. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment

standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 23: Summary Evaluation of PSU Tools–Access Security)

Facet 3: Item bank access security	
1. Describe the processes for access to the item bank.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Describe the security with respect to the physical access and computer installations.	G
b. Is there differentiated access with respect to the different user profiles? Could access be provided remotely? Why yes and why not?	F
c. Which is (are) the location(s) of the data file(s)?	E
d. What does the data file protection consist of (system file protection)?	A*
e. What backup procedures are there? How frequently are they implemented?	A*
f. Describe the auditing modules (log in / log out activities).	A*
g. Describe the types of outlets the system has: USB, print screen, access to the network, printing.	E

*Because there is no specific standard for Objective 1.1.d. Facet 3, no parenthetical listing of "Standards not met" will be made for a rating of "A." These ratings were based instead on the evaluators' understanding of commonly used practices in the assessment industry.

RECOMMENDATIONS

1. It is necessary to carry out a detailed inspection of the item bank system to determine the updating needs and modifications given the technological developments.
2. With respect to backups to the information systems, we recommend, if under further investigation it is found that the item bank is not supported with redundant systems,

a. to safeguard against interruptions in service and to maintain the most recent edits, establishing a redundant service (e.g., servers), and

b. to protect from catastrophic failure, scheduling incremental daily media backups and weekly full media backups of the database.

BIBLIOGRAPHY

DEMRE. (2011). *Banco de ítemes.* Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.d. Facet 4. Item bank – Process flow

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU item bank process flow.

There is no specific standard for this objective; reference is made to general recommendations on the standards oriented to safeguard the confidentiality of the information with issues related to the quality of the assessments, such as validity and reliability.

GENERAL DESCRIPTION

Three groups of people have access to the Item Bank:

- Information Technologies Unit
- Studies and Research Unit
- Test Construction Unit

Only this last one, the Test Construction Unit, has access to the questions in themselves, to their statements and the rest, and they are the only group that may carry out modifications. It is formed by six committees, one for each disciplinary field: Mathematics, Language and Communication, Social Sciences and History, and the three committees for Science (Biology, Chemistry and Physics). Each one of the six committees is formed by a head and its members. There also exists a head for the whole unit and a general coordinator.

After the outside (or inside) constructors have elaborated questions with the Safe Question software, these are recovered by each committee in accordance to the area it belongs to (the entering of the files with the items is done by means of USB) for which they require the respective password.

Each committee carries out the review of each item, and through consensus, its acceptance or rejection is decided upon. The reviews take place during working sessions. The head of each committee performs the loading of the approved items into the item bank platform. For that purpose, he/she registers the participants to the session where the item was reviewed and loads the statement and remaining data of the item. When loading each item, a single code is allotted to the item. It is possible for the codes loaded in the same session not to be consecutive, due to the fact that there may be several persons loading items.

If after being left in the bank the item is required for piloting, (for example), the commission reviews it again and, if it is necessary, performs changes on it that are generally very small and only formal. These changes take place in the platform with the use of Microsoft Word. For example, if deciding to include it in pre-testing, the item condition changes.

After the pre-testing, if the statistics indicate it is a good item, it is left under the condition of available. If the statistics are not good, then it is rejected. In this last case, the item is eliminated and is never tested again, or it may be used as an example for the population, or it is decided it is necessary to be redone, for which it must undergo the whole process again. That is, it returns to the commission and must be modified, etc.

The system maintains a register of all modifications done on the item during the whole process, be it a content modification or one on the correct response. The information of those participating is recorded, from which computer, the date, etc. During the whole process, the item keeps its single code.

An addition to the information mentioned previously, there is the very important information related to item statistics. Data processing through Test Classic Theory is done directly in the system, but the data processing by means of IRT is done with BILOG 3.11 software, outside the system, after which the data are re-entered into the system. This process is similar to what happens with the DIF processing done with DIFAS 4.0 software.

Even if its operation is not detailed, a document called Item Bank explicitly includes aspects on security, the item views, how to carry out test assembly, the results to the questions, their condition, the user privileges, etc. This manual is used only by members of the area committees.

EVALUATION

The description of the processes is extensive but not entirely clear. It is possible to perfect it and create the missing documentation, especially in that related to process manuals for the established procedures.

Table 24 shows a summary evaluation for PSU Tools (process flow). The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment

standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 24: Summary Evaluation of PSU Tools (Process Flow)

Facet 4: Process Flow (updating / adding records)	
1. Describe the item bank process flow (adding/updating records).	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. What reviews are carried out on the item bank administrator engine? Who is authorized to add/update the records in the item bank? • What type of quality control process is set in operation to assess a successful adding/updating process?	F
b. Which are the most important characteristics of your documentation process? • Which is the audience you have in mind with respect to the documentation process?	C**

**Because there is no specific standard for Objective 1.1.d. Facet 4, no parenthetical listing of "Standards not met" will be made for a rating of "C."

RECOMMENDATIONS

1. Currently, the item reviews take place based upon the expertise of the participants in the respective sessions (consensus is sought). However, there is no mention of manuals or standards to comply with. Therefore, we recommend that documenting the item review criteria.
2. Beyond the fact that the committee members seem to think that an item is good and possesses a certain difficulty, no other statistical or psychometric elements are applied. No statistical information analysis manuals are mentioned. Therefore, we recommend that committees additionally analyze items with respect to possible discrimination criteria, of the correct option or of the invalid options (distractors), or with an understanding that the items should function in some particular way.

BIBLIOGRAPHY

DEMRE. (2011). *Banco de ítemes.* Santiago: Universidad de Chile.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Objective 1.1.e. Quality of the terms used in the operative applications, considering the indicators used in their selection and considering indicators of item functioning (indicators of the Classical Test Theory, Item Response Theory, and DIF bias analysis) by genre, dependence and educational mode in the experimental sample and in the rendering population

Intended characteristics of items from pilot testing may not be present during operational administration of the items. Items contributing information to applicants' test scores are scrutinized for content irrelevant sources. Items deemed faulty may be removed from the computation of the total score due to issues raised on quality of the items. Statistical procedures available elsewhere may help in pinpointing items with potential issues. Additional human knowledge is also added to the statistical layer to compensate for statistical errors that may be present when analyzing a large number of items.

An example of an issue that handicaps an item is the existence of controversial answer keys. The evaluation team developed and performed interviews with relevant stakeholders from DEMRE on March 22, 2012. The interview process took a total of two hours from 11:45 to 13:45. The purpose of the interview was to gain a deeper understanding on:

- Quality criteria for judging items administered operationally (pilot samples of students and population of students) (Facet 1)
- Process for selecting operational items to render test scores (Facet 2)
- Process for reviewing and approving selected operational items (Facet 3)

All the interviews were performed within DEMRE offices following an agreed-upon schedule for the visit. The interview covered the three facets and relevant elements as agreed upon with the TC during the goal clarification meeting in Santiago, Chile, in January 2012.

The following DEMRE staff participated in the interviews:

- Head of the department of test construction
- Coordinator of test construction committees
    - Mathematics
    - Language
    - History and Social Studies
    - Science

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees.
The following subsections contain the results of the evaluation for Objective 1.1.e., Facets 1-3.

Objective 1.1.e. Facet 1. Criteria for pilot sample item and test taker population selection

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process to pilot sampling. A framework for evaluating PSU approaches for pilot sampling is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. (p. 44)

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures and evidence of model fit should be documented. (pp. 44-45)

GENERAL DESCRIPTION

The psychometric criteria used in the selection of items for the PSU are based in classical test theory (CTT) as well as from the item response theory (IRT).

The CTT criteria include difficulty, discrimination, and the distractor analyses. DEMRE has defined as acceptable a difficulty index of values falling within the range of $0.10 \leq p \leq 0.80$, which translates to $9.6 \leq \delta \leq 18.1$ when expressed on a delta scale.

According to the documentation reviewed, this range for the difficulty index varies in both limits for some tests, by virtue of previous analysis which have enabled to establish the low feasibility in obtaining items with very high or very low answer proportions (for example, the case of the Mathematics test, where it is infrequent to obtain very easy items). It is worthwhile noting that the precise reasons for establishing the criteria at these ranges were not found in the technical documentation nor in interviews with the DEMRE staff.

DEMRE has defined the acceptable discrimination index that is the equivalent of a minimum biserial correlation value of $r_b \geq 0.25$.

Concerning incorrect options (distractors), DEMRE has defined the following criteria:

- they must be elected by at least 2% or more of the applicants approaching the questions;
- they must present a negative biserial correlation coefficient (r_b);

- the average for the group elected by the distractor must be lower than the average for the group answering the correct answer and lower than the total group average.

Should one or several CTT indicators be out of range, "special conditions are established for item acceptance" (DEMRE, 2011, page 10ff), which is shown in the following table:

Table 25: DEMRE's Rules for Interpreting CTT Results

If a question evidences …

… an out of range difficulty percentage shall be valid if:

p_c	r_b
0.05 < p < 0.12	≥ 0.500
0.13 < p < 0.16	≥ 0.450
0.17 < p < 0.19	≥ 0.400

➡ p_d ≥ 0.03 in the remaining alternatives

p_c	r_b
0.91 > p > 0.88	≥ 0.300
0.87 > p > 0.84	≥ 0.275
0.83 > p > 0.80	≥ 0.250

➡ p_d ≥ 0.01 in the remaining alternatives

… voided distractors, shall be valid if:

p_d	6	p_c	r_b
0.010 – 0.014	< 11.5	> 0.66	≥ 0.300
0.015 – 0.019	< 13.5	> 0.46	
0.020 – 0.024	< 14.5	> 0.36	≥ 0.400
0.025 – 0.029	< 15.5	> 0.26	

p_d	6	p_c	r_b
0.010 – 0.014	< 12.4	> 0.56	≥ 0.400
0.015 – 0.019	< 14.4	> 0.36	
0.020 – 0.024	< 15.4	> 0.27	≥ 0.500
0.025 – 0.029	< 16.4	> 0.20	

Note: p_c= proportion of correct answers, r_b= biserial correlation coefficient, p_d= proportion of answers in distractors and ∆= delta difficulty scale.

The UCP has chosen the two parameter logistics model as the IRT model it will use for processing the PSU tests data. In that model, the following indicators are taken:

- The ICC (Item Characteristic Curve)
- The *b* difficulty value must be between $-5 \leq b \leq 5$, a range which happens to be more extensive than that accepted in international environments, and item discrimination at ≥ 0.6 (DEMRE, 2011a, page 27)
- Item and test information function

The following table shows the interpretation values of the IRT discrimination parameter: (DEMRE, 2011b, page 19).

Table 26: DEMRE's Rules for Interpreting IRT Discrimination Results

a parameter range values	Item discrimination classification
≤ 0.0	Does not discriminate
0.01 – 0.34	Very low discrimination
0.35 – 0.64	Low discrimination
0.65 – 1.34	Moderate discrimination
1.35 – 1.69	High discrimination
≥ 1.70	Very high discrimination

(DEMRE, 2011b)

The criterion used for the PSU to use items with discrimination is the parameter range equal to or greater than the classification of moderate discrimination by the table above. Note, however, that the item discrimination criterion found on page 27 (an acceptable range of discrimination being greater than or equal to 0.60) does not align with DEMRE's rules as articulated in Table 26, which includes 0.60 to 0.64 in the low range of discrimination.

Another indicator considered and used in item selection is the DIF, calculated through the Mantel-Haenszel method and it is delivered to the Breslow-Day statistic. "When through these statistics it is proven that the item possesses a differential functioning, DEMRE provides a series of estimators indicating the magnitude of these differences and if the item favors the focus group or the referential group" (DEMRE, 2011b, page 24/25).[4]

[4] Please note the original quotation: "Cuando se comprueba a través de estos estadísticos que el ítem posee un funcionamiento diferencial, el DEMRE entrega una serie de estimadores que indican la magnitud de estas diferencias y si el ítem favorece al grupo focal o referencial" (DEMRE, 2011b, page 24/25).

In accordance with the qualitative classification realized by the ETS of the DIF magnitude, there are three DIF categories: Irrelevant (A), Moderate (B) and Severe (C). In PSU item selection, irrelevant DIF items are included preferably and, if necessary, the decision is taken to use some items classified as Moderate (B). Items cataloged as Severe DIF (C) are never used.

These indicators are described in the technical documentation that the test assemblers have available, ensuring by this that they are informed about the same and that they know the acceptance and non-acceptance criteria of each indicator. Since the item selection for assembly takes place directly in the item bank's information system and since the indicators for each item appear in these technical files, the professional in charge of assembly has this information available at the moment of carrying out the item selection. Neither the technical documentation nor the interviews with DEMRE staff elicited information with respect to the procedure (e.g., identifying a hierarchy) for making decisions when an item complies with some criteria established but not others. That is, there is no evidence of the establishment of criterion priority or a specific protocol for decision making on this issue in particular.

The current version of the item selection protocol corresponds to an "updating of the psychometric guidelines regarding item acceptance or elimination delivered by the Research and Studies Unit (UEI) in 2005 to the UCP, for its application to the items after their experimentation"[5] (DEMRE, 2011b). That is, a review of the criteria that had been applied since 2005 took place in 2011.

EVALUATION

It is evident that, in general, DEMRE uses clear criteria on item selection, viz., indicators of the classic theory and IRT (2 parameters). In almost all of those criteria, the established acceptance criteria corresponds to internationally accepted ranges, with the exception of the IRT difficulty indicator, which provides results that are quite more extensive than commonly accepted, as well as the omission level, which happens to be quite elevated, even though it is not the same for all tests. The differentiation in the criteria applied for different tests is explained. However, it is not supported by studies evidencing that such differences have no effects upon the assessment process, taking into account the purpose of the test, the assessed population and the object of assessment. In addition, there is no documentation that describes either basis for the criteria used for item selection or a procedure to be followed for selecting an item when some criteria are met and other criteria are not met.

Table 27 shows a summary evaluation for the PSU Criteria towards pilot sample item and test taker population selection. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

[5] Original quotation: ". . . actualización de las directrices psicométricas para la aceptación o eliminación de ítemes entregado por la Unidad de Estudios e Investigación (UEI) en el año 2005 a la UCP, para ser aplicadas a los ítemes después de su experimentación" (DEMRE, 2011b, nota al pie, página 2).

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 27: Summary Evaluation of PSU Criteria towards Pilot Sample Item and Test Taker Population Selection

Facet 1: Criteria towards pilot sample item and test taker population selection	
1. Describe the quality criteria regarding operationally administered items (Pilot sample and test taker population).	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Explain the criteria followed for the assessment of the quality regarding the operationally administered items (psychometric, contents, purpose/use). • Explain the process for exploring the item quality of the operationally administered throughout the subpopulations of interest (for example, gender, curricular branch, type of high school, region, SES). • Are these criteria consistent among the pilot sample and test taker populations?	E
b. What are the bases (clarity and rationality) for the criteria?	D*
c. Which is the hierarchy or preference for the different criteria	D*

types (psychometric vs. content)? • What relevant policy on fairness and equality is utilized? Which is the process to review/modify the policy on fairness and equality?	

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. We recommend documenting the reason for the criteria transformation in item selection between 2005 and 2011; specifically, documenting the test history in its technical procedures and the reasons for performing changes on the same.
2. We recommend specifying how the IRT indicators are analyzed: ICC and Information Function. Are they the criteria for acceptance or rejection?
3. Although psychometric criteria for the item difficulty established in general correspond to internationally accepted criteria -3 to +3, the evidence from our demonstration of the IRT processes for the PSU test showed a range of ability closer to the range -5 and +5. In order to judge which of these two ranges is more appropriate for test assembly, we recommend to documenting how many questions exist with difficulty levels that are closer to and further away from the extremes of this range -5 and +5.
4. We recommend reviewing and reconciling the differing criteria for acceptable IRT discrimination values (i.e., $a \geq 0.6$ versus $a \geq 0.65$), given that items selected with that criterion would be cataloged at a low discrimination level, in accordance with the classification table for this indicator.
5. We recommend reviewing the current criterion used for flagging items for high omission rates, applying a standard drawn from the observations and experience of the evaluation team with respect international assessments (i.e., 10% omissions).

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2011a). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2011b). *Directrices psicométricas para el análisis de ítems PSU.* Santiago: Universidad de Chile.

ETS. (2002). *Standards for quality and fairness*. Princeton N. J. Author.

IEA. (1999). *Technical standards for IEA studies*. Eburon Academic Publishers. Netherlands. Author.

Objective 1.1.e. Facet 2. Item selection process — Pilot sample and test taker population

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process for selecting items. A framework for evaluating the PSU approaches for selecting items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. (p. 44)

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures and evidence of model fit should be documented. (pp. 44-45)

GENERAL DESCRIPTION

The item selection for test assembly is the responsibility of the head of each committee, along with other members determined by the head.

> The committees are formed by three to five members, each one of which presents degrees and academic specialization types encompassing adjoining areas of each subsector:
>
> - In Language and Communication: linguistics, literature, curriculum and assessment.
> - In Mathematics: numbers, proportionality, algebra, functions, geometry, probabilities and statistics, curriculum and assessment.
> - In History and Social Sciences: Chilean, American and universal history, curriculum and assessment.
> - In Science: Biology, Physics and Chemistry, including the different specialties proper to each theme axis by discipline. (DEMRE, 2011a, p.23)

The main tools for selecting items designed for test assembly are the test specifications tables and the technical criteria for item acceptance for official tests.

Through the specifications tables, the person in charge of item selection determines the number of items per table cell that must be selected per each test section, discipline area or theme axis, as well as by the cognitive skills assessed and required question format, in accordance with the proportion designated in the same specifications tables. That is, the item selection abides by to two types of criteria: quantitative (the number of items to be selected per each portion of the specifications table) and qualitative (compliance with the pre-established psychometric criteria for each one of the indicators used).

Once the head of the respective committee assigns one of the fellow professionals the task of performing the assembly of a test, the necessary authorizations and passwords are activated, in order for that professional to be able to start the item selection process. The professional proceeds with a first selection based upon the quantitative and qualitative criteria, printing the files for those items for review by a peer, who is generally another member of the committee.

The process combines automatic processes with manual processes. Since the items are loaded in the item bank system, searches occur automatically according to the item condition and criteria references that have been entered into the database; however, the final selection also requires a personal decision that rests with the expert eye of the professional responsible for test item selection. That person must decide which items to include and which not to include from among the several items complying with the requested criteria. This is particularly evident in the item selection regarding the Language test, since "the texts residing in the Item Bank may include questions associated in different storage levels": "approved by the commission," "approved by the pre-test," "rejected by the pre-test," "available for assembly," etc. The different storage levels make managing the process of text assembly and its questions particularly difficult. There must be not only an adequate item distribution in accordance with class and type (note Section 3 of the current Language test, context vocabulary and reading comprehension questions associated to a text), but also an adequate distribution per theme axes, statistical results and storage condition in the item bank (DEMRE, 2011a, p.25).

In order to carry out an adequate selection process, there are several stages of analysis that must be performed. These stages are:

- qualitative and quantitative analysis of the pilot sample application results of the previous year;
- tracing of item classes, according to the test section (Language);
- tracing of text types and associated questions (Language);
- item classification according to theme axis;
- item classification according to cognitive skill;
- item classification according to statistical difficulty;
- item classification according to discrimination; and, eventually,
- DIF detection, that is, checking to see if the item presents a bias.

"Once the previous phases are completed, the assembly procedure begins properly such" (DEMRE, 2011a, p.22).

It is worth noting that, in addition to the criteria stated regarding item selection,

> the committees are concerned with analyzing the omission as additional background information, operating as a reference to the purpose of selection of the test being assembled. (DEMRE, 2011a, p.27)

It must be pointed out that there exists a general agreement regarding the acceptable criteria of the different psychometric indicators used in item selection. Variations are applied concerning some of these indicators in consideration of particular characteristics of the population taking the test or of the contents of same. These particularities are described following:

> In the case of the Language test, the preference is to work with the omission with a maximum range of 30%, since a very small group of applicants reaches the maximum score with the correct solution to the 80 questions of the instrument.
>
> Dealing with a test which is much more oriented towards measuring skills and competencies, it is not necessary either to include too many questions with a high degree of difficulty being the acceptable range in the order of 10% (Δ 18.1) to 90% (Δ 7.9).
>
> In the case of the remaining tests, the difficulty fluctuates around 40%, with the exception of the History test, which is assembled with 47%. In general, the items are ordered according to theme axes, by difficulty level, from the easiest to the most difficult. In the case of History, the chronological order is taken into consideration, in the geographical field, from the general to the particular, always considering the subject matter.
>
> In Mathematics, the difficulty of each question fluctuates between 3% and 80%, and its omission could reach, in some questions, 70% eventually, with the evident purpose of optimizing discrimination among the population segment with better results.
>
> Regarding the other tests, the omission of some questions could reach even 60%, and the difficulty of each question fluctuates between 7% and 80%. (DEMRE, 2011a, p. 29)

When an item shows extreme behavior or falls outside the criteria of one or more of the indicators used, the item is analyzed by the respective committee responsible for weighing the statistical and non-statistical information of the item in order to achieve consensus about whether to accept or reject the item.

The rejected items are used as instructional material during the construction process to exemplify for the item constructing commissions the typical construction mistakes and to explain the interpretation of the item psychometric indicators.

EVALUATION

The decisions on item selection for operational tests are shared among the members of the technical team who, as it has been described, have academic and psychometric qualifications that enable them to perform this labor. The decisions are based upon team review stages and discussions that confer reliability to the process. In addition to the general criteria listed in here, more specific criteria for operationalizing the selection of items for operational tests are described above in Objective 1.1.c. Facet 6. In general, the process appears to be adequate and meets with minimal expectations.

However, there is a need for more precise documentation that would describe in a systematic way whether priority should be given to certain item indicators over others when deciding between items to be included in a test. It would be useful if the software used in item construction allowed developed items to be loaded to the bank with the history of their modifications and uses in particular administrations. This would allow, for example, easy access to items rejected in the pilot that could be useful as training material for the item writers.

The procedures for dealing with items showing extreme behaviors are acceptable because such items are, in fact, reviewed by the appropriate DEMRE content committee using relevant criteria.

Table 28 shows a summary evaluation for the PSU item selection process (Pilot sample and test taker population). The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment

standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 28: Summary Evaluation of PSU Item Selection Process (Pilot Sample and Test Taker Population)

Facet 2: Item selection process (Pilot sample and test taker population)	
1. Describe the process for operational item selection.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Which is the plan with respect to the selection process (when, who, etc.)?	D*
b. How are the selection criteria operationalized?	E
c. What are the qualifications of the personnel selecting the items?	E
d. What procedures exist with respect to items with extreme behaviors (for example, atypical)?	E
e. Is the item selection process manual or automatic?	E

***A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. We recommend documenting in greater detail the item selection process regarding what steps are entailed in its planning and how specific criteria are applied to each test. Additionally, in cases where there is partial fulfillment of the psychometric criteria on the part of an item, we recommend documenting the rationale that determines which indicator is to have priority in front of the rest.
2. We recommend modifying the software used in item construction so that developed items could be loaded to the bank with the history of their modifications and uses in administration.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2005). *Criterios de análisis de las preguntas de una prueba experimental*. Santiago: Universidad de Chile.

DEMRE. (2011a). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

DEMRE. (2011b). *Protocolo de análisis de ítemes, rendición oficial y asignación de puntajes PSU.* Santiago: Universidad de Chile.

Objective 1.1.e. Facet 3. Review and approval of selected operational items (Pilot sample and test taker population)

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process to review and approve operational items. A framework for evaluating the PSU approaches for reviewing and approving operational items is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.7

> The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. (p. 44)

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures and evidence of model fit should be documented. (pp. 44-45)

GENERAL DESCRIPTION

The review and approval of the selected operational items takes place through meetings of the committees responsible for the tests, who analyze the statistical information of the items, previously loaded in the item bank system after the application of the tests.

> The quality criteria regarding the item review process are authenticated by five (5) concurrent proceedings:
>
> - The curricular reference file, which is represented in the item elaboration software
> - The guideline for item review
> - The review, comment, discussion, adequacy and, eventually, the reformulation of the items submitted in the commission
> - The review, comment, discussion, adequacy and, eventually, the reformulation of the items carried out by the President of the Commission and the Technical Advisor, before assembling a pilot test
> - The review, comment, discussion, adequacy and, eventually, the reformulation of the items carried out in the corresponding committee, before assembling a pilot test (DEMRE, 2011, p.19)

The results of the item review are recorded in the Item Review Guideline form, which includes the following item description data:

- Name identification of the constructor
- Identification of the session number and date it took place
- Identification of the item class (regarding Language, test section and subsection; concerning other tests, this box indicates if the item is a direct, combined or sufficient data question)
- Item designation ID (DEMRE, 2011, p.20)

The Item Review Guideline also includes the data that serve the purpose of properly assessing the item quality, such as:

- Indicators used in the review
- EM level ID: identification of the high school level to which the item is referred to
- Theme axis ID: identification of the curriculum theme axis of the study area to which the item is referred to
- CMO ID: identification of the minimum mandatory contents of the curriculum of the area to which the item is referred to
- Study Program Content ID: identification of the study program content corresponding to the respective level of the Language curriculum to which the item is referred to
- Cognitive Skill ID: identification of the cognitive skill defined in the Curricular Reference Matrix of the respective area to which the item is referred to
- Focus ID: identification of the pedagogical focus to which the item is referred to, in the case of Language
- Difficulty ID: identification of the estimated difficulty to which the item is referred to
- Pertinence ID: identification of the item's curricular pertinence (Can we evaluate this CMO? Is it possible to evaluate such content study program? Is the cognitive ability consistent with the CMO, content and mode of preparation of the item?)
- Relevance ID: identification of the curricular and assessment relevance of the item (Is the item representative of the main thematic axis, the CMO, the study program, or the cognitive skill to be assessed?).

- Format ID: identification of the formal aspects of the item: according to test section (Language, Science) and according to the kind of question (Language, Mathematics)
- Syntax ID: identification of the adequate form of organizing the statement, stimulus or question forming the item
- Correct Option Key ID: identification of one single answer option that solves the problem posed by the item
- Distractor ID: identification of the homogeneity and quality of the distractors accompanying the item correct answer
- Final grading: the expression of one of the five possible levels for item grading (DEMRE, 2011, p.21)

All of these descriptors are the object of contrast with the previously established criteria for the definition of the acceptance or rejection of the items. In the Language Test Commission, the indicators in the preceding paragraph are valued in accordance with the following scale:

- 4. Optimum: the item reaches the excellence level. The item may be entered to the item bank under the 'proposed" condition, so that later, when edited, it can pass into the condition of "approved by the commission."
- 3. Good: the item does not reach the superior condition, but it is elaborated well. The item may be entered in to the Item Bank under the condition of "proposed," so that later, when edited, it can pass into the condition of 'approved by the commission."
- 2. Sufficient: the item presents inadequacies between two (2) and four (4) fields of the assessment guideline. The item must be modified for its approval, or on the contrary it shall be entered into the item bank under the condition of "rejected."
- 1. Rejected: the item is not approved by the commission, and it is entered into the item bank under the condition of "rejected by the commission." (DEMRE, 2011, p.20)

In the case of the remaining commissions, the final decision for item assessment is reported in two categories only: approved or rejected.

In general, the item review process counts with the participation, in different moments, of the following personnel:

- The head of the respective DEMRE committee
- The members of the committee
- The members of the item writer commission in which the Technical Advisor and the President of the Commission are included

- The outside reviewer, in his capacity as an academician at the Universidad de Chile

Interviews with DEMRE indicated that the participants in this process are professionals educated in the areas evaluated; many of them are said to have postgraduate education in assessment and measurement. In this sense, it may be said that the personnel involved in item review and approval have the required qualifications to carry out these processes in accordance with the guidelines provided to them.

EVALUATION

The teams responsible for item selection for operational tests rely on documentation for the item indicators and characteristics that must be taken into account. However, DEMRE does not have an established procedure for making item selections when an item complies with some criteria but not others. This allows for greater subjectivity in selecting items and may affect the comparability of forms across administrations.

There is no evidence in the reviewed documentation that the question selection procedures include comparisons between the behavior of the items in pilot and operational applications.

The *Optimum/Good/Acceptable/Rejected* scale presented above explicitly allows piloted items to be edited or otherwise changed prior to operational use. This acceptance scale contradicts best practices in operational test form development. Items should not be edited or changed unless the items are re-piloted.

Table 29 shows a summary evaluation for the PSU Selected Operational Item Review and approval (pilot sample and test taker population). The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 29: Summary Evaluation of PSU Selected Operational Item Review and Approval (Pilot Sample and Test Taker Population)

Facet 3: Selected Operational Item Review and Approval (pilot sample and test taker population)	
1. Describe the process for operational item review and approval.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Which is the process with respect to reviews and approval of the items to be elected for their operational administration (when, who, etc.)?	E
b. How are the review and approval processes and the results documented? Which is the process for managing the statistical flags with respect to the items elected for their operational administration?	E
c. What indicators are used in the operational item review process?	F
d. What indicators are used in the operational item approval process?	C (3.9)
e. Which are the qualifications of the personnel involved in the review and approval processes? How much personnel training takes place using the test construction specifications?	E
f. What do you do when the statistics of the administered items collapse between the pilot and operational administrations? • Policy around the decision? Please point out the main components and the key actors formulating the policies. • Any research around the decisions? Please point out the main avenues towards research and discoveries.	C (3.9)

RECOMMENDATIONS

1. We recommend documenting the reasons that were considered for establishing some differences in the review of items from the Language committee and the rest of the committees. We also recommend analyzing the possibility of standardizing these proceedings for all tests, in as much as possible. Whenever this is not possible, we recommend that the arguments be documented and the controls be anticipated so that the differences do not affect the results or otherwise be counterproductive with respect to the purpose of the test.

2. We also recommend that over time the choice of participants for the review processes be deliberately made to increase institutional and geographical diversity. Finally, we recommend documenting the policies with respect to contingencies, such as the number of items not approved by the established criteria being higher than is regularly found.
3. We strongly recommend that piloted items not be edited or otherwise changed prior to operational use unless the items are re-piloted.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2011). *Descriptores técnicos del proceso de construcción de instrumentos de evaluación para las pruebas de selección universitaria.* Santiago: Universidad de Chile.

Objective 1.1.f. Degree of consistency between the indicators of item functioning obtained in the application on the experimental sample regarding those obtained in the rendering population

The international evaluation team relied on professional standards for appraising the merit and worth of PSU item functioning. A framework for evaluating the PSU approaches for item functioning is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and / or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented. (pp. 44-45)

DESCRIPTION OF PROCESS
Purpose of the objective

The evaluation team is studying factors associated to differential performance of items between their pilot and operational administrations. Several factors are known to cause item statistics to drift between administrations, and their contributions to the explanation of change in item performance are studied utilizing measures of association (the Pearson *r*-coefficients). Analyses are performed for each PSU test in the battery separately. Only items with pilot and operational performance are considered for the analyses. The purpose of the analyses is to summarize patterns of consistency of item performance (percent of omissions, CTT, IRT and DIF) on pilot and operational administrations. Analyses are disaggregated to study generalization of patterns on the following subpopulations of interest:

- Gender
- Curricular branch (Scientific-Humanistic and Technical-Professional)
- Type of school (Municipal, Subsidized and Private)
- Socioeconomic status (defined with data from the family income variable)
- Region (Metropolitan, North and South)

This analysis will allow us to determine subpopulation variables correlated to greatest variability in item statistics from pre-test to operational use. This information can then be used to infer changes to pre-testing strategies, or operational item selection strategies that might be beneficial to the PSU program.

Analysis performed

The study intends to respond to the following question:

- What is the degree of consistency of the item statistics between the pilot application and the official assembly of the PSU tests, among the following subpopulations?
    - Gender
    - Curricular branch (Scientific-Humanistic and Technical-Professional)
    - Type of school (Municipal, Subsidized and Private)
    - Region (Metropolitan, North and South)

These analyses were carried out regarding each test according to the information provided. The tests are identified in the tables of results by means of a code, which may be read in the following table:

Table 30: Code of the Analyzed Tests

CODE	TEST
LC	Language and Communication
M	Mathematics
HSS	History and Social Sciences
S	Science
S-Com	Science – Common Module
S-ElecBio	Science – Elective Biology Module
S-ElecPh	Science - Elective Physics Module
S-ElecCh	Science - Elective Chemistry Module
S-ComBio	Science – Common Biology Module
S-ComPh	Science – Common Physics Module
S-ComCh	Science – Common Chemistry Module

Report

Several factors which may affect item performance between piloting and the official test assembly are studied. The contribution of each factor to statistical change between test applications is studied by means of the Pearson *r*-coefficient. The analyses are carried out for each PSU test in each year of official application. Only the items for which statistics have been obtained in piloting as well as the official application are considered. The purpose of the analysis is to summarize the item performance consistency patterns in the pilot and official administrations of the tests.

The analysis is done on the statistics: percentage of omissions, CTT difficulty, IRT difficulty and discrimination and CTT discrimination (biserial correlation). Note: on some occasions the biserial correlation is called biserial or biserial correlation maintaining the original name appearing on the databases. The analysis in the following populations is disaggregated in order to study the pattern generalizations:

- Gender
- Curricular branch (Scientific-Humanistic and Technical-Professional)
- Type of school (Municipal, Subsidized and Private)
- Region (Metropolitan, North and South)

Outstanding information:

- From 2009 onward, the information from the Science area has been grouped, which makes the analysis lose detail.
- The lower coherence is found in the data of the biserial correlation and especially concerning the cases of the type of school (dependency) and in a few cases of gender.
- It is interesting to observe that in several occasions the estimation for IRT difficulty shows low associations between piloting and the official administration of the test, since greater stability would have been expected due to the characteristics of the model.
- The lower association values are present in the Language and Communication test (LC); in some years, it is quite low.

General data analysis

For the analyses below, a total of 5102 items with complete information from the pilot, as well as from the final administration, were studied and their performance summarized in figures, associations, and, when feasible, effect sizes.[6]

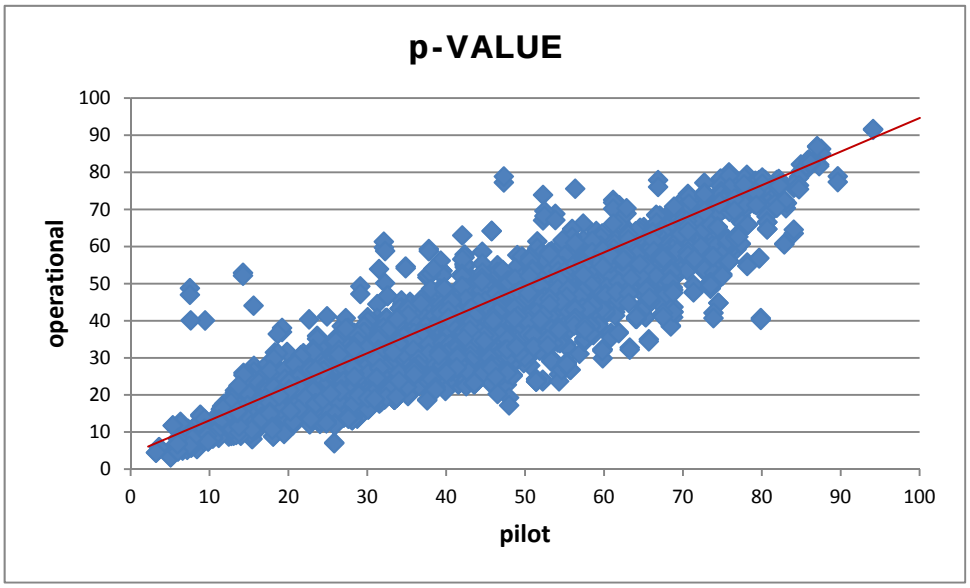


Figure 1: Association between the CTT Difficulty Values from the Pilot and Operational Administrations

The difficulty values are expressed in a scale from 0 to 100. In accordance with the results, it is clear that the relation is linear and that there is a tendency in the sense that the values of the pilot administration are greater than those of the operational administration. As well,

[6] The graphs submitted in this section correspond to an approximation; it is technically incorrect to carry out the comparison if they are not equated, though the exercise aids in detecting outliers.

a few cases are observed where the values of the operational administration are very different and greater than those of the pilot administration.

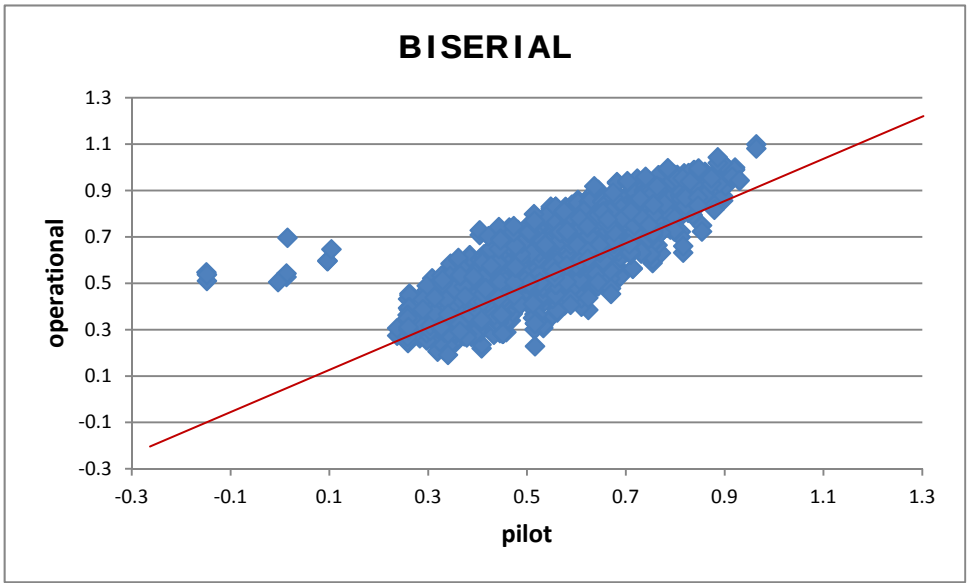


Figure 2: Association between the Biserial Correlation Values from the Pilot and Operational Administrations

The relation between both administrations is linear. Because the items are the ones that have been administered operationally, their biserial correlation values are positive in the pilot as well as in the operational administration, that is, they already went through an item analysis. Generally, the values are slightly higher in the operational administration. A few data points are observed outside the general trend with much higher values in the operational administration. It is worthwhile noting that a few questions have biserial correlation values above 1.0.

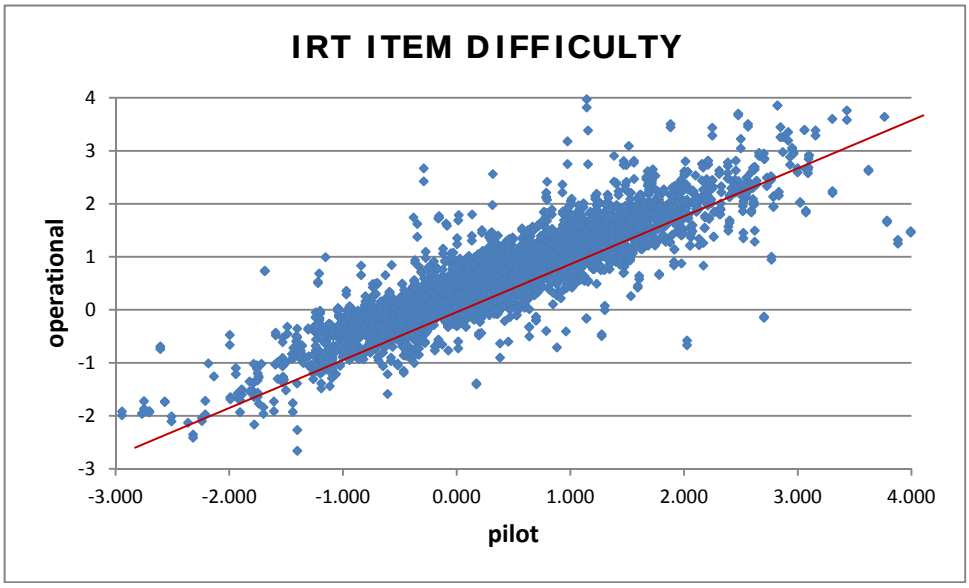


Figure 3: Association between the IRT Difficulty Values from the Pilot and Operational Administrations

In general terms, the values of the operational administration are greater than in the pilot administration. The graph does not show 3 items that have difficulty values above 4.0 (including one with a value of 12). The number of items with very different values between one administration and the other is greater than those of the CTT difficulty indices. Considering that one of the assumptions of the IRT is the conservation of the independent values of the sample, it is not clear why such large differences are observed in this graph (in general 0.5 logit and in some cases, differences above 1.0).

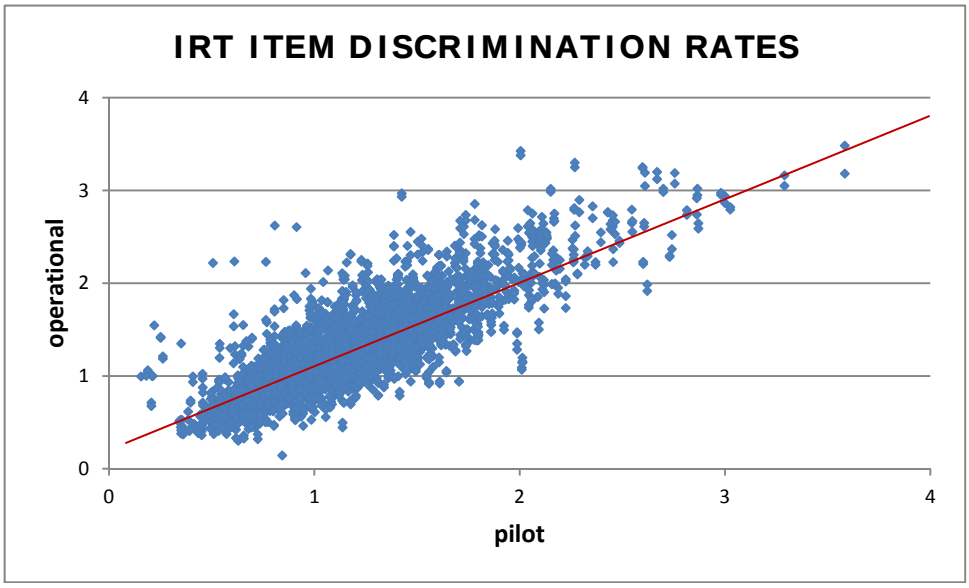


Figure 4: Association between the IRT Discrimination Values from the Pilot and Operational Administrations

Most of the values are located between 0 and 2. The relation is linear, and in general, the values are larger in the operational administration.

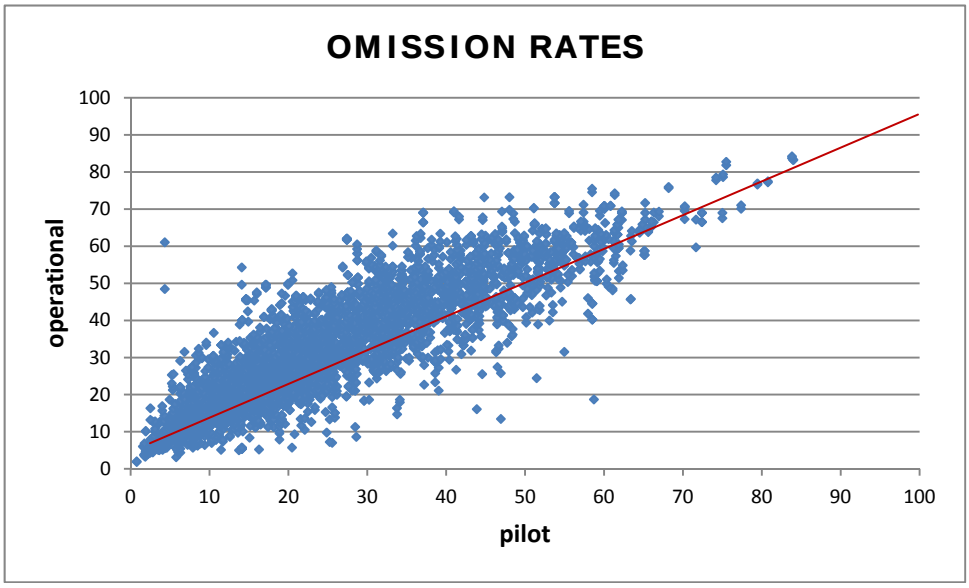


Figure 5: Association between the Omission Rate Values from the Pilot and Operational Administrations

The large difference in omission rate values between the pilot and operational administrations may be explained by the test takers' knowledge that a correction for guessing formula is applied in scoring the operational administration. There are more than 800 items with omission rates greater than 50% in the operational administration (i.e., around 15% of the items). These omission values have an important influence over the estimation of the item parameters and indices, which may explain the differences observed in Figure 1 through Figure 4.

Table 31. Maximum Value, Minimum Value and Quartiles for Differences between Pilot and Operational Statistics

	P-VALUE	BISERIAL	IRT DIFFICULTY	IRT DISCRIMINATION	OMISSION
Minimum Value	-41.260	-0.696	-12.422	-1.815	-56.730
Quartile 1	0.590	-0.113	-0.507	-0.289	-11.520
Quartile 2	4.960	-0.063	-0.256	-0.115	-6.460
Quartile 3	9.710	-0.014	-0.006	0.023	-2.520
Maximum Value	39.660	0.288	12.874	0.942	40.040

The evaluation team calculated the difference in values between the pilot and operational administrations for each of the indicators as shown in Figure 1 through Figure 5. This yielded the maximum and minimum values of these differences as well as the quartiles.

More than 75% of the pilot items have p-values greater than the p-values for the operational administration. This shows that, in general, the items are easier when piloted than when included on an operational form. This finding is corroborated by the results in the third column for IRT difficulty, where at least 75% of the pilot items have values less than for the operational items.

With respect to the values of biserials, over 75% of the items have somewhat higher values when piloted than during the operational administration. This indicates that the piloted items are somewhat more discriminating than when used operationally. The IRT discrimination parameter differences indicate much the same results as the biserial differences.

Finally, there is a higher percentage of omissions in the pilot (a little over 75% of cases) than for the operational administration. Note that the differences in the rates of omission between the pilot and operational administrations may have an impact on the difficulty of the items across administrations.

Together, all of these results show that there are differences in the behavior of the items between pilot and operational administrations that are due to differences in the samples of students or in their item-answering behavior.

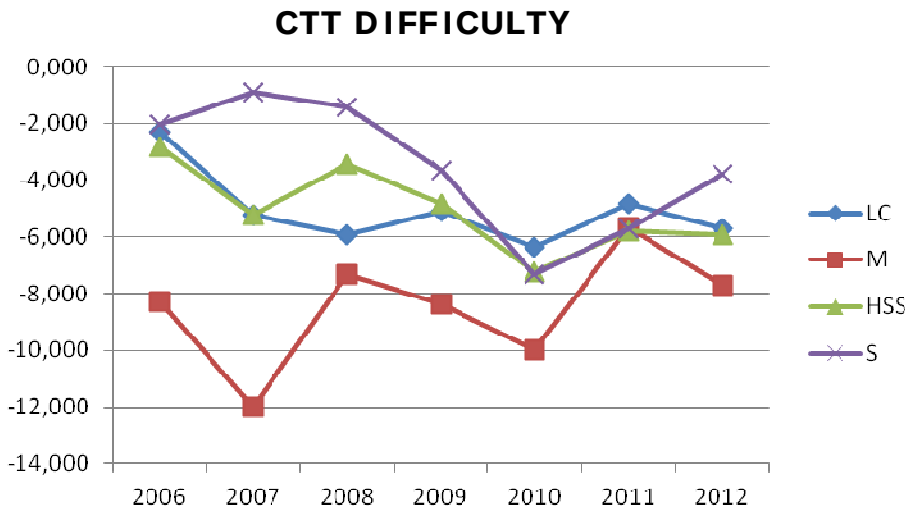


Figure 6: Plot of the Medians of the Differences between Operational and Pilot CTT Difficulties across Year by PSU Test

These data capture the similarity between the statistics derived from the operational and pilot administrations over time. When the closer the medians are to zero, the greater the similarity between the central tendencies of the operational and pilot distributions of statistics.

It is observed in the figure that the medians of the differences are not stable between 2006 and 2012 for the different tests. For example, there is a great variation between the medians of differences for the Science and Mathematics tests for 2007. Yet there is very little variation between them in 2011. There are no clear trends beyond the fact that the medians are closer to each other starting in 2009.

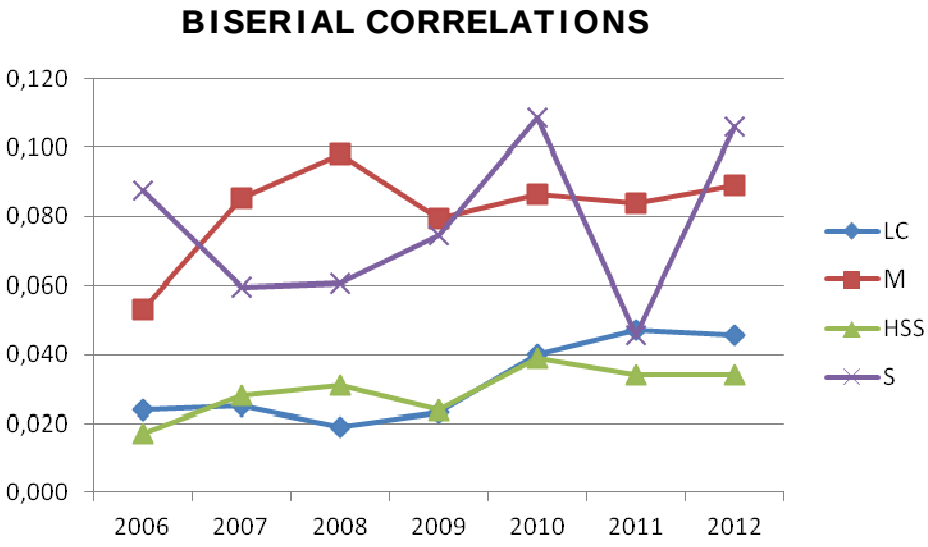


Figure 7: Plot of the Medians of the Differences between Operational and Pilot CTT Biserials across Year by PSU Test

The trends of the History and Social Sciences test and the Language and Communication test are well-behaved; the values of the median differences do not vary greatly across the years. The data of the Language and Communication test vary during the first three years and later become stable. The Mathematics data show the largest median differences. The stability of the majority of the tests indicates that the sets of piloted items and operational items are very similar, producing similar results in both cases.

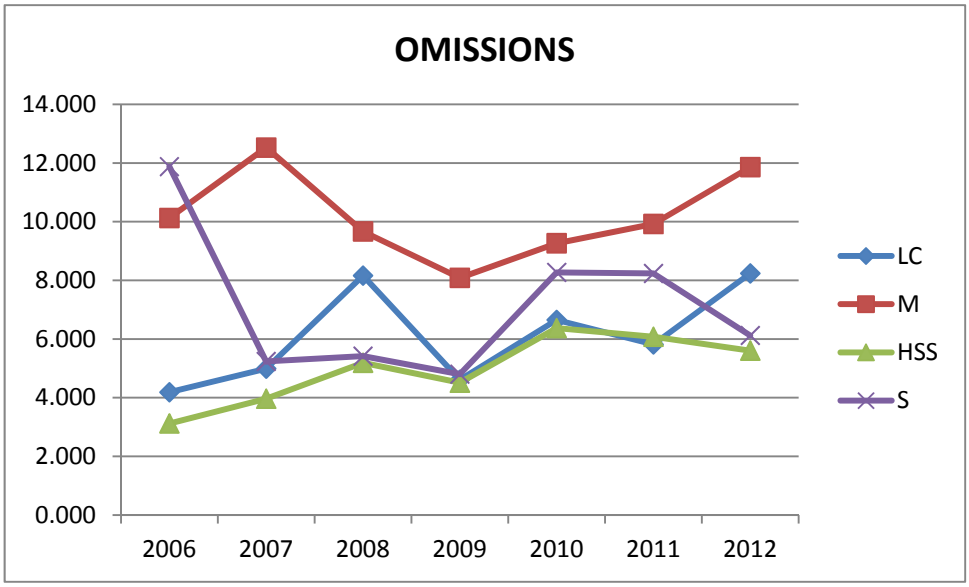


Figure 8: Plot of the Median Differences of the Omission Values across Year by PSU Test

Starting from 2007, the behavior of three of the tests is relatively similar and stable. That is, the omission rates in the pilot administration are similar to the operational administration. Mathematics shows the greatest median differences, even though the tendency is relatively stable for all years; these high values in the median of the differences indicate that the sample behavior is very different from that of the operational population: the omissions rate in the pilot administration is lower than in the operational administration.

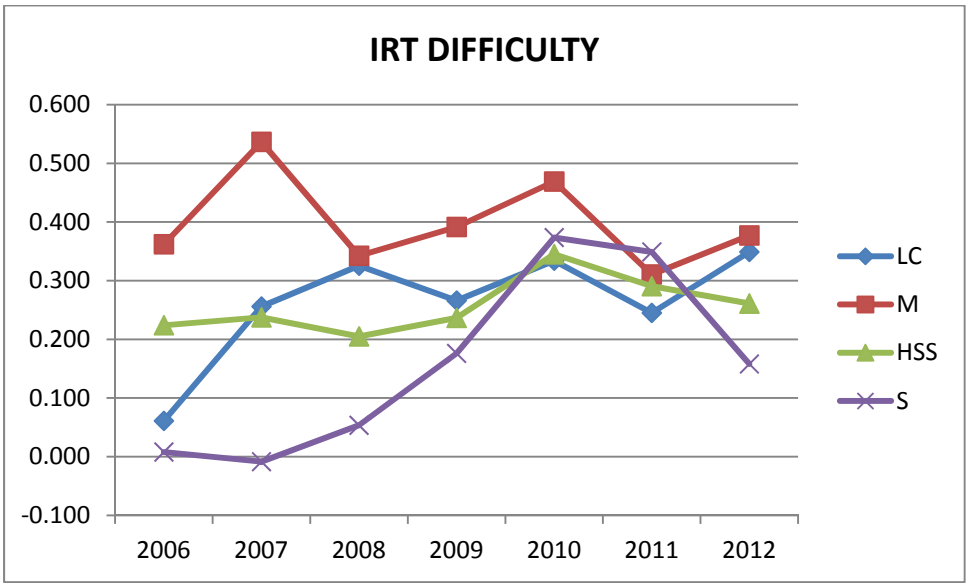


Figure 9: Plot of the Median Differences of the IRT Difficulty Values across Year by PSU Test

Because IRT values are not on the same scale from year to year in any subject and between pilot and operational administrations, caution should be exercised when interpreting these plots.

In general, the IRT difficulties seem to be higher for items when they appear on the operational form than when they appear on the pilot form. Starting in 2007, the data for the History and Social Sciences test and Language and Communication test are stable and very similar. In Science, the trend was for the IRT difficulty to increase during the first years, but to stabilize in 2010 and 2011, and to decrease in 2012. The greatest difference in the median of the IRT difficulties is shown in Mathematics. In almost all cases from 2010 until 2012, the median of the differences surpasses one fourth logit (0.25), even reaching 0.4 logit, indicating that the behavior of the items in the pilot administration is different from that of the operational administration.

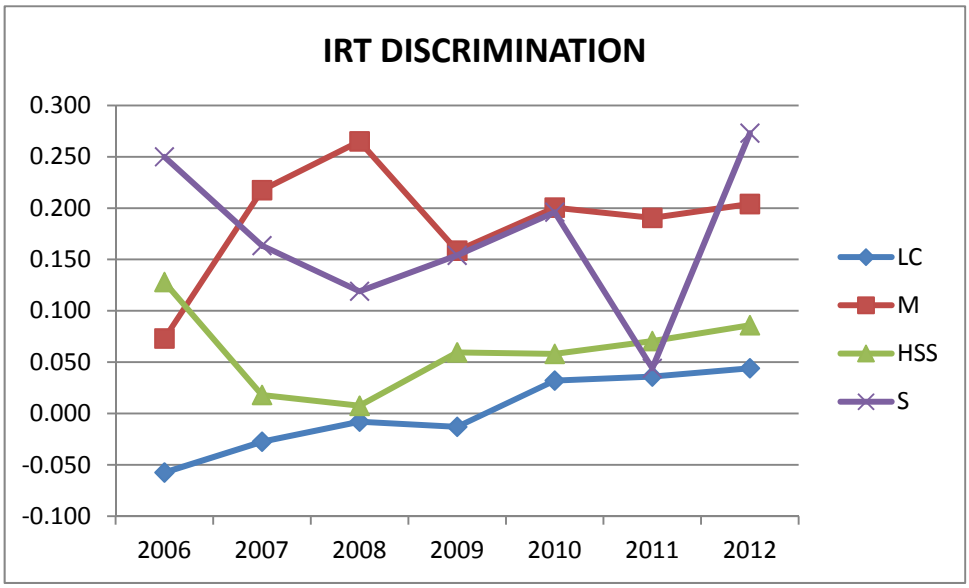


Figure 10: Plot of the Median Differences of the IRT Discrimination Values across Year by PSU Test

Once again, the same caveat with respect to any interpretations regarding IRT values should be exercised here.

For most years and subjects, the IRT discrimination parameters seem to be higher for items when they appear on the operational form than when they appear on the pilot form, except for Language and Communication. Such as in previous cases, the more stable tendencies are those of History and Social Sciences and Language and Communication. Mathematics is the most unstable, with the most variable tendency. In this case, Science also shows a large variation and instability.

2006

In general terms, the values for gender and type of school show low levels of association between piloting and the official application. Similarly, the value is low regarding the Technical-Professional curricular branch for Science - Elective Chemistry Module (S-ElecCh).

In the case of biserial correlation, the lowest levels of association between piloting and the official application are those of the Technical-Professional curricular branch, especially for Science – Elective Biology (S-ElecBio), whose value is negative. Some specific cases of low values are those of the Metropolitan Region in Language and Communication and the Science-Elective Chemistry Module and those of the daylight Scientific-Humanistic (HC) curricular branch in Language and Communication and the Science-Elective Chemistry Module.

The lowest values for the association, for all tests, are those of gender and dependence.

Even though a certain association is expected for values calculated by IRT, attention is drawn due to the low values for Language and Communication and History and Social Sciences regarding difficulty and History and Social Sciences concerning discrimination.

Table 32: 2006—Number of Items per Test between Piloting and Official Assembly

ITEM #	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio
GENERAL	160	130	142	38	42	52	2
REGION	140	88	90	38	42	42	2
GENDER	160	102	140	38	42	52	2
DEPENDENCE	160	102	140	38	42	52	2
CURRICULAR BRANCH	140	86	86	38	42	42	2
IRT	160	100	131	38	42	52	2

Table 33: 2006—Association between Difficulty Values

DIFFICULTY	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh
GENERAL	0.889	0.893	0.890	0.963	0.857	0.862
METROPOLITAN REGION	0.874	0.954	0.860	0.958	0.853	0.895
OTHER REGIONS	0.898	0.956	0.917	0.960	0.854	0.848
MALE	0.394	0.468	0.334	0.488	0.372	0.303
FEMALE	0.429	0.512	0.342	0.546	0.331	0.274
PRIVATE	0.380	0.532	0.354	0.412	0.349	0.309
SUBSIDIZED	0.411	0.508	0.312	0.542	0.354	0.298
MUNICIPAL	0.421	0.476	0.343	0.505	0.386	0.222
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.890	0.973	0.922	0.969	0.854	0.873
TECHNICAL-PROFESSIONAL	0.907	0.922	0.882	0.921	0.815	0.645

Note that the values of association are much lower for gender and dependency for all tests in 2006. This may be due to the lack of representativeness or sample size of the pilot in these cases.

Table 34: 2006—Association between Biserial Correlation Values

BISERIAL	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh
GENERAL	0.650	0.852	0.865	0.875	0.898	0.834
METROPOLITAN REGION	0.484	0.888	0.736	0.816	0.827	0.644
OTHER REGIONS	0.702	0.825	0.880	0.802	0.827	0.763
MALE	0.975	0.990	0.990	0.979	0.994	0.983
FEMALE	0.973	0.990	0.989	0.984	0.983	0.986
PRIVATE	0.988	0.989	0.982	0.971	0.987	0.985
SUBSIDIZED	0.973	0.982	0.987	0.98	0.987	0.984
MUNICIPAL	0.978	0.976	0.986	0.976	0.974	0.971
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.619	0.919	0.897	0.848	0.874	0.766
TECHNICAL-PROFESSIONAL	0.519	0.668	0.621	-0.165	0.448	0.585

Table 35: 2006—Association between Omission Values

OMISSION	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh
GENERAL	0.826	0.873	0.900	0.922	0.875	0.871
METROPOLITAN REGION	0.789	0.931	0.838	0.926	0.874	0.877
OTHER REGIONS	0.827	0.936	0.902	0.91	0.865	0.841
MALE	0.295	0.486	0.321	0.415	0.374	0.319
FEMALE	0.269	0.518	0.332	0.443	0.296	0.302
PRIVATE	0.226	0.473	0.293	0.325	0.353	0.286
SUBSIDIZED	0.295	0.513	0.325	0.452	0.366	0.304
MUNICIPAL	0.282	0.521	0.336	0.442	0.338	0.327
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.803	0.947	0.909	0.922	0.864	0.869
TECHNICAL-PROFESSIONAL	0.796	0.933	0.880	0.884	0.828	0.732

The associations for gender and dependency are very low, which may be indicating that the pilot sample is not representative or is too small in these cases.

Table 36: 2006—Association between IRT Values

IRT	LC	M	HSS	S-ElecBio	S-ElecCh	S-ElecPh
DIFFICULTY	0.440	0.924	0.405	0.885	0.839	0.825
DISCRIMINATION	0.687	0.912	0.302	0.876	0.928	0.826

Given the IRT characteristics and assumptions, such low levels of association for LC and HSS would not be expected. This could be related to the size of the pilot samples.

2007

In general, low levels of association occur especially in the biserial correlation and in relation to gender and dependency. In the case of the variable region and the tests of the common Science module, there are very few items, which is why calculations of the association were not carried out.

In the case of the associations for the difficulty value, in no case are the values lower than 0.7, indicating a high consistency between the pilot data and the official application.

Two negative associations for the biserial correlation are observed in the Technical-Professional branch.

Table 37: 2007—Number of Items per Test between Piloting and Official Assembly

N	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	160	128	116	50	46	52	22	26	24
REGION	160	96	70	50	46	50	4		2
GENDER	160	106	114	50	46	52	22	26	24
DEPENDENCE	160	106	114	50	46	52	22	26	24
BRANCH	160	96	66	50	46	52	22	26	24
IRT	160	106	109	50	46	50	22	26	26

Table 38: 2007—Association between Difficulty Values

DIFFICULTY	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	0.957	0.871	0.850	0.938	0.719	0.907	0.958	0.897	0.954
METROPOLITAN REGION	0.950	0.929	0.775	0.936	0.936	0.936	N/A	N/A	N/A
OTHER REGIONS	0.954	0.926	0.845	0.934	0.934	0.934	N/A	N/A	N/A
MALE	0.948	0.892	0.850	0.917	0.763	0.896	0.946	0.898	0.955
FEMALE	0.958	0.906	0.862	0.945	0.418	0.898	0.943	0.906	0.951
PRIVATE	0.939	0.934	0.856	0.928	0.802	0.830	0.958	0.900	0.956
SUBSIDIZED	0.949	0.899	0.851	0.927	0.672	0.909	0.939	0.899	0.949
MUNICIPAL	0.959	0.901	0.843	0.915	0.759	0.906	0.958	0.925	0.948
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.958	0.946	0.821	0.938	0.700	0.902	0.970	0.899	0.899
TECHNICAL-PROFESSIONAL	0.931	0.909	0.803	0.886	0.818	0.790	0.835	0.937	0.937

Table 39: 2007—Association between Biserial Correlation Values

BISERIAL	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	0.864	0.778	0.740	0.776	0.810	0.835	0.924	0.959	0.93
METROPOLITAN REGION	0.813	0.762	0.528	0.732	0.732	0.732	N/A	N/A	N/A
OTHER REGIONS	0.839	0.841	0.706	0.746	0.746	0.746	N/A	N/A	N/A
MALE	0.782	0.846	0.680	0.717	0.834	0.809	0.871	0.970	0.940
FEMALE	0.855	0.825	0.709	0.694	0.486	0.784	0.968	0.939	0.883
PRIVATE	0.663	0.737	0.681	0.659	0.790	0.771	0.897	0.907	0.806
SUBSIDIZED	0.827	0.829	0.689	0.675	0.509	0.803	0.878	0.959	0.899
MUNICIPAL	0.798	0.768	0.624	0.641	0.550	0.711	0.874	0.893	0.908
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.850	0.893	0.752	0.735	0.765	0.836	0.936	0.953	0.953
TECHNICAL-PROFESSIONAL	0.366	0.73	0.201	-0.091	-0.056	0.405	0.361	0.572	0.572

The associations for the Technical-Professional curricular branch are very low. This may be an indication of a lack of measurement precision for this subpopulation.

Table 40: 2007—Association between Omission Values

OMISSION	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	0.930	0.901	0.873	0.913	0.707	0.842	0.962	0.904	0.882
METROPOLITAN REGION	0.927	0.933	0.714	0.892	0.892	0.892	N/A	N/A	N/A
OTHER REGIONS	0.920	0.928	0.800	0.921	0.921	0.921	N/A	N/A	N/A
MALE	0.930	0.926	0.859	0.906	0.731	0.862	0.939	0.878	0.889
FEMALE	0.920	0.927	0.868	0.91	0.581	0.825	0.952	0.926	0.877
PRIVATE	0.918	0.950	0.850	0.913	0.809	0.862	0.957	0.948	0.939
SUBSIDIZED	0.931	0.935	0.863	0.910	0.700	0.815	0.958	0.894	0.876
MUNICIPAL	0.914	0.932	0.857	0.88	0.648	0.819	0.959	0.916	0.848
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.928	0.954	0.781	0.913	0.725	0.846	0.976	0.906	0.906
TECHNICAL-PROFESSIONAL	0.897	0.921	0.786	0.859	0.696	0.771	0.911	0.907	0.907

Table 41: 2007—Association between IRT Values

IRT	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
DIFFICULTY	0.933	0.928	0.796	0.92	0.858	0.889	0.936	0.947	0.917
DISCRIMINATION	0.839	0.843	0.762	0.765	0.914	0.798	0.861	0.968	0.905

2008

In general terms, the coherence of the values between the pilot and the official application is high. However, there is a particular association achieved for the Technical-Professional branch, which, as seen in Table 44, has a biserial correlation of 0.007 for Science – Elective Chemistry Module, whose value indicates no relationship at this rate in the two administrations (pilot and final). The presence of such a low association may be an artifact of sample size caused by low enrollment of students from Technical-Professional backgrounds on pilot test administrations and operational administrations of the PSU Chemistry test.

Table 42: 2008—Number of Items per Test between Piloting and Official Assembly

N	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	160	134	148	52	52	52	34	36	36
REGION	160	128	146	52	44	52	20	24	24
GENDER	160	132	148	52	52	52	34	36	36
DEPENDENCE	160	132	148	52	52	52	34	36	36
BRANCH	160	126	146	52	52	52	34	36	34
IRT	160	130	148	50	44	50	14	26	20

Table 43: 2008—Association between Difficulty Values

DIFFICULTY	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	0.896	0.929	0.922	0.933	0.913	0.947	0.949	0.948	0.969
METROPOLITAN REGION	0.893	0.924	0.922	0.919	0.916	0.951	0.962	0.965	0.973
OTHER REGIONS	0.889	0.921	0.901	0.929	0.905	0.937	0.949	0.944	0.957
MALE	0.880	0.929	0.920	0.918	0.923	0.921	0.949	0.976	0.967
FEMALE	0.899	0.931	0.916	0.930	0.880	0.951	0.947	0.924	0.968
PRIVATE	0.854	0.944	0.906	0.891	0.898	0.868	0.933	0.941	0.963
SUBSIDIZED	0.896	0.940	0.907	0.928	0.907	0.958	0.948	0.941	0.970
MUNICIPAL	0.892	0.922	0.913	0.909	0.934	0.944	0.936	0.972	0.977
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.886	0.940	0.917	0.935	0.892	0.946	0.963	0.935	0.935
TECHNICAL-PROFESSIONAL	0.837	0.896	0.893	0.874	0.975	0.824	0.877	0.964	0.964

Table 44: 2008—Association between Biserial Correlation Values

BISERIAL	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	0.823	0.844	0.830	0.752	0.841	0.664	0.900	0.760	0.739
METROPOLITAN REGION	0.747	0.824	0.741	0.748	0.794	0.484	0.887	0.437	0.825
OTHER REGIONS	0.774	0.800	0.775	0.647	0.841	0.670	0.846	0.611	0.897
MALE	0.772	0.821	0.699	0.659	0.857	0.747	0.904	0.713	0.619
FEMALE	0.781	0.822	0.803	0.677	0.605	0.481	0.812	0.769	0.821
PRIVATE	0.675	0.648	0.498	0.402	0.802	0.374	0.847	0.505	0.763
SUBSIDIZED	0.732	0.78	0.775	0.489	0.725	0.653	0.848	0.665	0.535
MUNICIPAL	0.722	0.828	0.701	0.549	0.818	0.623	0.779	0.730	0.637
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.805	0.854	0.790	0.706	0.819	0.659	0.859	0.747	0.747
TECHNICAL-PROFESSIONAL	0.524	0.507	0.587	0.414	0.007	0.444	0.729	0.641	0.641

The association for S-ElecCh for the professional technical branch is almost null; i.e., the values of the biserial correlation between the pilot and the operational do not resemble each other. The values of the association are low for this branch in general. Once again it is worthwhile to analyze the sample selection and the qualification rules as possible causes for this fact.

Table 45: 2008—Association between Omission Values

OMISSION	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
GENERAL	0.822	0.908	0.882	0.846	0.726	0.830	0.922	0.868	0.864
METROPOLITAN REGION	0.803	0.899	0.893	0.823	0.715	0.820	0.730	0.929	0.944
OTHER REGIONS	0.823	0.893	0.861	0.855	0.622	0.834	0.806	0.901	0.893
MALE	0.816	0.905	0.895	0.828	0.747	0.838	0.919	0.930	0.850
FEMALE	0.815	0.906	0.869	0.849	0.690	0.824	0.916	0.796	0.871
PRIVATE	0.792	0.916	0.879	0.846	0.812	0.804	0.892	0.796	0.889
SUBSIDIZED	0.813	0.919	0.886	0.833	0.721	0.827	0.931	0.881	0.873
MUNICIPAL	0.786	0.901	0.852	0.844	0.657	0.806	0.910	0.866	0.842
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.821	0.905	0.891	0.854	0.766	0.825	0.931	0.863	0.863
TECHNICAL-PROFESSIONAL	0.705	0.915	0.833	0.843	0.681	0.711	0.851	0.871	0.871

Table 46: 2008—Association between IRT Values

IRT	LC	M	HSS	S-ElecBio	S-ElecPh	S-ElecCh	S-ComBio	S-ComPh	S-ComCh
DIFFICULTY	0.853	0.933	0.919	0.896	0.954	0.921	0.958	0.914	0.982
DISCRIMINATION	0.864	0.835	0.842	0.854	0.836	0.643	0.944	0.802	0.781

2009

From 2009 onward, the information on the Science tests has been grouped into one test. No information is found by region. The consistency is high for difficulty values.

Consistently, low values of association are found in the biserial correlation for all cases of the Language and Communication. The same happens with the IRT difficulty.

Table 47: 2009—Number of Items per Test between Piloting and Official Assembly

N	LC	M	HSS	S
GENERAL	160	138	149	478
REGION	160	138	149	478
GENDER	160	138	149	478
DEPENDENCE	160	134	138	466
BRANCH	160	138	148	386

Table 48: 2009—Association between Difficulty Values

DIFFICULTY	LC	M	HSS	S
GENERAL	0.857	0.927	0.908	0.892
REGION	0.848	0.927	0.902	0.910
GENDER	0.860	0.921	0.905	0.865
DEPENDENCE	0.749	0.942	0.907	0.905
BRANCH	0.868	0.934	0.897	0.898
IRT	0.879	0.923	0.903	0.866
GENERAL	0.845	0.951	0.926	0.904
REGION	0.888	0.953	0.873	0.855

Table 49: 2009—Association between Biserial Correlation Values

BISERIAL	LC	M	HSS	S
GENERAL	0.563	0.817	0.799	0.865
REGION	0.530	0.835	0.783	0.800
GENDER	0.559	0.800	0.755	0.833
DEPENDENCE	0.519	0.561	0.560	0.701
BRANCH	0.555	0.774	0.747	0.764
IRT	0.553	0.703	0.682	0.803
GENERAL	0.550	0.821	0.819	0.863
REGION	0.496	0.756	0.596	0.338

For the LC test the association values are a little low for all cases of the sample. This fact may be caused by the selection of the sample for this test.

Table 50: 2009—Association between Omission Values

OMISSION	LC	M	HSS	S
GENERAL	0.896	0.912	0.893	0.840
REGION	0.890	0.919	0.891	0.853
GENDER	0.890	0.902	0.889	0.816
DEPENDENCE	0.900	0.935	0.894	0.854
BRANCH	0.893	0.914	0.889	0.847
IRT	0.865	0.918	0.879	0.803
GENERAL	0.887	0.930	0.915	0.850
REGION	0.830	0.927	0.857	0.808

Table 51: 2009—Association between IRT Values

IRT	LC	M	HSS	S
DIFFICULTY	0.389	0.943	0.898	0.894
DISCRIMINATION	0.752	0.892	0.826	0.809

Such as in the biserial correlation, the sample effect is also observed here in the low association of LC in the difficulty values.

2010

There is no information on gender and region variables. In general, the association values are high with the exception of the biserial correlation regarding LC, some cases of Private dependence and the difficulty value concerning LC of the IRT.

Table 52: 2010—Number of Items per Test between Piloting and Official Assembly

N	LC	M	HSS	S
GENERAL	160	140	147	480
GENDER	160	138	147	
DEPENDENCE	160	138	147	480
BRANCH	160	102	141	
IRT	160	138	147	456

Table 53: 2010—Association between Difficulty Values

DIFFICULTY	LC	M	HSS	S
GENERAL	0.930	0.944	0.925	0.871
REGION	0.850	0.945	0.927	
GENDER	0.866	0.941	0.919	
DEPENDENCE	0.786	0.952	0.918	0.863
BRANCH	0.866	0.958	0.911	0.872
IRT	0.868	0.902	0.899	0.851
GENERAL	0.852	0.960	0.933	
REGION	0.885	0.947	0.876	

Table 54: 2010—Association between Biserial Correlation Values

BISERIAL	LC	M	HSS	S
GENERAL	0.642	0.838	0.791	0.812
METROPOLITAN REGION	0.326	0.757	0.749	
OTHER REGIONS	0.381	0.881	0.747	
MALE	0.414	0.478	0.478	0.614
FEMALE	0.354	0.752	0.673	0.691
PRIVATE	0.325	0.784	0.766	0.773
SUBSIDIZED	0.397	0.831	0.799	
MUNICIPAL	0.291	0.632	0.676	

As in the past year, here low associations are also observed in the LC test. Once again, this fact may be due to the sample selection.

Table 55: 2010—Association between Omission Values

OMISSION	LC	M	HSS	S
GENERAL	0.907	0.931	0.880	0.755
METROPOLITAN REGION	0.893	0.935	0.877	
OTHER REGIONS	0.911	0.932	0.874	
MALE GENDER	0.914	0.955	0.865	0.795
FEMALE GENDER	0.913	0.952	0.859	0.763
PPAG DEPENDENCE	0.865	0.883	0.869	0.722

Table 56: 2010—Association between IRT Values

IRT	LC	M	HSS	S
DIFFICULTY	0.547	0.964	0.922	0.849
DISCRIMINATION	0.659	0.844	0.828	0.799

2011

The association values between pilot and official application concerning the biserial correlation are low for Language and Communication and some cases of History and Social Studies; the same regarding PPAG (Private) dependence.

Table 57: 2011—Number of Items per Test between Piloting and Official Assembly

N	LC	M	HSS	S
GENERAL	160	136	148	478
REGION	124	98	96	198
GENDER	160	140	148	478
DEPENDENCE	160	140	148	478
BRANCH	160	130	146	476
IRT	160	134	148	472

Table 58: 2011—Association between Difficulty Values

DIFFICULTY	LC	M	HSS	S
GENERAL	0.851	0.889	0.933	0.835
METROPOLITAN REGION	0.811	0.922	0.923	0.764
OTHER REGIONS	0.821	0.899	0.950	0.683
MALE	0.835	0.879	0.922	0.841
FEMALE	0.858	0.899	0.934	0.837
PRIVATE	0.787	0.888	0.909	0.827
SUBSIDIZED	0.859	0.905	0.945	0.851
MUNICIPAL	0.836	0.820	0.886	0.793
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.851	0.909	0.929	0.847
TECHNICAL-PROFESSIONAL	0.821	0.702	0.919	0.839

Table 59: 2011—Association between Biserial Correlation Values

BISERIAL	LC	M	HSS	S
GENERAL	0.670	0.846	0.795	0.828
METROPOLITAN REGION	0.541	0.773	0.699	0.731
OTHER REGIONS	0.589	0.806	0.643	0.737
MALE	0.639	0.804	0.770	0.801
FEMALE	0.613	0.829	0.750	0.700
PRIVATE	0.544	0.580	0.501	0.603
SUBSIDIZED	0.687	0.811	0.714	0.749
MUNICIPAL	0.595	0.722	0.628	0.684
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.626	0.857	0.795	0.820
TECHNICAL-PROFESSIONAL	0.514	0.440	0.577	0.401

Once again low values appear for the LC test, even though, in general, all the values are lower than in the previous years.

Table 60: 2011—Association between Omission Values

OMISSION	LC	M	HSS	S
GENERAL	0.870	0.912	0.897	0.797
METROPOLITAN REGION	0.868	0.952	0.880	0.798
OTHER REGIONS	0.880	0.944	0.907	0.767
MALE	0.860	0.902	0.905	0.807
FEMALE	0.859	0.916	0.881	0.783
PRIVATE	0.794	0.866	0.882	0.777
SUBSIDIZED	0.877	0.931	0.901	0.802
MUNICIPAL	0.823	0.869	0.867	0.770
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.874	0.918	0.902	0.803
TECHNICAL-PROFESSIONAL	0.771	0.648	0.878	0.741

Table 61: 2011—Association between IRT Values

IRT	LC	M	HSS	S
DIFFICULTY	0.797	0.914	0.885	0.855
DISCRIMINATION	0.745	0.879	0.792	0.792

2012

The coherence values regarding the biserial correlation values are low for most of the dependence cases.

Table 62: 2012—Number of Items per Test between Piloting and Official Assembly

N	LC	M	HSS	S
GENERAL	156	148	150	474
REGION	116	30	52	150
GENDER	156	146	150	474
DEPENDENCE	156	146	150	474
BRANCH	156	142	148	474
IRT	156	146	150	456

Table 63: 2012—Association between Difficulty Values

DIFFICULTY	LC	M	HSS	S
GENERAL	0.948	0.922	0.937	0.898
METROPOLITAN REGION	0.945	0.960	0.936	0.914
OTHER REGIONS	0.943	0.944	0.942	0.936
MALE	0.936	0.920	0.916	0.903
FEMALE	0.952	0.932	0.943	0.873
PRIVATE	0.928	0.945	0.870	0.859
SUBSIDIZED	0.945	0.929	0.941	0.898
MUNICIPAL	0.922	0.891	0.907	0.866
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.940	0.959	0.936	0.897
TECHNICAL-PROFESSIONAL	0.926	0.929	0.921	0.865

Table 64: 2012—Association between Biserial Correlation Values

BISERIAL	LC	M	HSS	S
GENERAL	0.816	0.855	0.847	0.760
METROPOLITAN REGION	0.675	0.460	0.742	0.807
OTHER REGIONS	0.760	0.725	0.597	0.762
MALE	0.750	0.822	0.782	0.766
FEMALE	0.783	0.812	0.835	0.670
PRIVATE	0.585	0.475	0.707	0.584
SUBSIDIZED	0.797	0.843	0.737	0.626
MUNICIPAL	0.721	0.687	0.662	0.685
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.771	0.861	0.830	0.752
TECHNICAL-PROFESSIONAL	0.559	0.750	0.523	0.267

Table 65: 2012—Association between Omission Values

OMISSION	LC	M	HSS	S
GENERAL	0.835	0.912	0.931	0.839
METROPOLITAN REGION	0.827	0.929	0.914	0.834
OTHER REGIONS	0.843	0.926	0.914	0.839
MALE	0.799	0.903	0.916	0.819
FEMALE	0.853	0.923	0.930	0.830
PRIVATE	0.865	0.945	0.900	0.832
SUBSIDIZED	0.853	0.925	0.937	0.848
MUNICIPAL	0.764	0.890	0.900	0.787
DAYLIGHT SCIENTIFIC-HUMANISTIC	0.840	0.948	0.938	0.849
TECHNICAL-PROFESSIONAL	0.761	0.920	0.892	0.741

Table 66: 2012—Association between IRT Values

IRT	LC	M	HSS	S
DIFFICULTY	0.925	0.960	0.936	0.885
DISCRIMINATION	0.859	0.865	0.821	0.746

EVALUATION

The figures with the mean information of the differences between the values of the pilot administration and the operational administration show that it is necessary to restate once again the issue of the sample selection especially for the Mathematics and Science tests, which are the ones presenting the least stable tendencies and greatest variations in almost all parameters and indices. This issue also has to be analyzed in detail for the other two tests. It seems clear that the behavior of those assessed in the sample is very different from those assessed in the operational administration. It is worthwhile reflecting on whether the pilot administration is pertinent, given the fact that the motivation of those assessed is different from that of those who take the operational test form, which imposes the system of correction for guessing.

The fact of taking the tests in their subdivisions (Science) leads to a relatively low number of items for performing comparisons in each one of the years analyzed, in some cases reaching an insufficient number for performing comparisons. This takes place within 2006 and 2009 for the Science test.

Based upon results of analyses performed with both pilot and operational item performance across years (see Figure 1 through Figure 5 and Table 31), there are significant differences in the item performance indicators between the pilot administration and the operational administration. Though the relation is positive, the amount of variance accounted for is low for particular combinations of the PSU test and the subpopulations of interest.

Considering the results of the classical test theory values as a whole (difficulty, biserial correlation and omission) it is found that the item behavior is not the same in the pilot as in the final administration. The differences are large and significant when the item values are analyzed for the set of all of the years in all tests. For example, the category of omissions increases substantially (7%) in the final administration; the assumption is that those assessed, upon learning that the final qualification uses a correction formula, omit those responses for which they have a reasonable doubt. This fact affects, in itself, in a great

measure the values of all statistic estimations, causing the item values (not only those of the CTT, but also those of the IRT) to be underestimated or overestimated.

This issue is of no lesser interest since the number of questions omitted by international educational assessments is not as high as that reached by the PSU assessments. According to the *PISA 2006 Technical Report* (Organisation for Economic Co-Operation and Development, 2009, p. 219), the weighted average of omitted items was 5.41. In the *PISA 2009 Technical Report* (Organisation for Economic Co-Operation and Development, 2012, p. 200) the average number of omitted items, 4.64, was slighting smaller than in 2006. This suggests that a reasonable upper limit for omissions might be more in the order of 10%. These values, as may be observed, are very low with respect to the 25% and 32% recorded average in the pilot and final administrations of the PSU, respectively.

The same effect occurs with the IRT values, which are different between both test administrations. This contradicts the theoretical underpinning of IRT that these values are independent from the sample they are obtained from, if they are representative of one same population.

For all years, the lowest association values between the pilot and the final administration correspond to the biserial correlation: specifically, for gender, dependency and branch. It is necessary to note that these discrimination values are affected by the total score obtained from all test items and the strategies used to respond. That is, the discrimination values are affected by the items with technical problems or by how the pilot is managed, or even by the fact that students know about the correction for guessing used to score the test. In this regard it is expected that the values of the biserial correlation between the pilot and the final administration change substantially, which is seen in the tables and figures. It may also be the case that biserial computations can be affected by difficulty levels of items and thus become low as a result of that methodological artifact.

A high level of omission rates and the use of correction for guessing may also have contributed to discrepancies on item performance between pilot and operational administrations. Because the correction for guessing scoring is known to the applicants, the CTT pilot indices may not be reliable approximations of item performance on the operational context.

In summary, the pilot does provide important information about the quality of the items that can be used to make decisions about them in terms of their inclusion or not in the bank for use in any final administration. However, it is not clear that the data can be used as precise values of the different indices studied; i.e., changes of the values obtained in the pilot and the final administration are more or less large and significant. These changes occur mainly due to the lack of consideration of all variables (gender, for example) in the sampling of the population for the pilot administration; or, in some cases, the lack of participants forces the sampling plan (the branch, for example) to be reconstructed.

Secondly, the effect of the correction for guessing system of scoring may induce different strategies among students when participating in the pilot and final administrations. The analyses performed on rates of omission showed larger rates of omission for operational administration than for pilot administrations.

Thirdly, the fact that the pilot is a voluntary administration modifies the selected sample. Table 67 shows a summary evaluation for PSU Quality, security and confidentiality standards regarding the development of items and tests. The purpose of the table is to

provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 67: Summary Evaluation of Degree of Consistency between the Indicators of Item Functioning Obtained in the Application on the Experimental Sample regarding those Obtained in the Rendering Population

Facet 1: Degree of consistency between the indicators of item functioning obtained in the application on the experimental sample regarding those obtained in the rendering population	
1. Establish a comparison between the values of the items in the pilot and the final administration	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Is the information of the indicators of item functioning clear and usable?	E

RECOMMENDATIONS

1. We recommend taking steps so that the changes of the values obtained in the pilot and the final administration are closer together. For example, greater consideration of variables such as gender should be made during the sampling of the population for the pilot administration.
2. In accordance with the previous recommendation, and other evidence in the evaluation, it is recommended that DEMRE reconsider the use of formula scoring in the context of the PSU. Such formula scoring is based on theoretical assumptions with weak support and international university admission programs have abandoned its use or are seriously considering removing it from their processes.
3. We recommend analyzing the impact of rates of non-participation on intended representation of major socio-demographic variables during the pilot sampling process.
4. We recommend redefining the elements of the sample design of the pilot administration, taking into account the purpose of said administration, the purpose of the PSU, the psychometric theory to be used in the item and test analysis and the scoring scale to be used. This redefinition includes considering other forms of item piloting such as the inclusion of item groups in the operational administration. These groups of items would not be scored or used to obtain the results of those answering them, but that would provide statistics very close to the data from the operational administrations since those being assessed would not know which items are being piloted.
5. Although the sample participating in the pilot is voluntary and there is a commitment from the institutions selected to have their students participate, it is important to analyze impact of non-participation on intended representation of major socio-demographic groups in such a way as to account for possible bias in the results. Once this analysis is preformed, the historical non-participation rates could be accounted for with over-sampling of those groups going forward. For example, Levy and Lemeshow (1999) describe methods for dealing with non-response and missing data in sample surveys.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications (3rd edition)*. New York: John Wiley & Sons.

Organisation for Economic Co-Operation and Development. (2009). *PISA 2006 technical report.* OECD Publishing. Retrieved from: http://www.oecd.org/pisa/pisaproducts/pisa2006/42025182.pdf

Organisation for Economic Co-Operation and Development. (2012). *PISA 2009 technical report.* OECD Publishing. Retrieved from: http://www.oecd.org/pisa/pisaproducts/pisa2009/50036771.pdf

Objective 1.1.g. Exploration of variables associated to DIF, in case it is present

The exploration of variables related to PSU differential item functioning (DIF) comprises three layers of analyses. The first one involves expert analyses of DIF processes documented in DEMRE reports and clarified during interviews of DEMRE staff. The second layer involves analytical inspection of variables related with DIF. The inspection covered statistical modeling of DIF results with relevant item level information. The third layer provides results of an empirical demonstration of DIF computation with PSU data from the 2012 admission process. This layer of empirical analyses responded to a need to explore DIF for subpopulations that historically have not been covered by DIF analyses such as region (North, South and Central), socio-economic status (quintiles from family income and parents' level of education) and type of curriculum (Scientific-Humanistic and Technical-Professional). Historically, DEMRE has performed DIF explorations for subpopulations based on gender and dependency (Municipal, Subsidized and Private).

The report is organized according to the above three layers. The expert analyses are summarized following a structure that is similar to previously presented objectives. In this structure, a general description of DEMRE processes is presented previous to expert evaluation, summary, and recommendations. See Facet 1.

Facet 2 covers the analytical inspection of variables associated to DIF. It reports prevalence of DIF on pilot and operational administrations and a plausible locus of origin for DIF. Items with both pilot and operational DIF information are considered for the inspection, and analyses are performed for each PSU test in the battery separately. Databases cover the admission process from 2006 to 2012. For earlier PSU administrations, DIF analyses were not part of DEMRE's psychometric quality control processes. Our analyses are directed at (1) summarizing DIF magnitude and intensity for pilot and operational administrations by PSU test and year and (2) modeling logit (Mantel-Haenszel DFI flag vs. no flag), using logistic regression (LR) with the following set of predictors: (1) CTT and IRT difficulty and discrimination, (2) percent of omissions, (3) gender, (5) type of school (Municipal, Private and Subsidized).

Finally, DIF demonstration, the last portion of the report for Objective 1.1.g., covers DIF summary results with PSU data from the 2012 test administration and main subpopulation of interest. The analyses are geared to investigate differential item functioning for a comprehensive set of subpopulations: Socio-economic status (SES), Region (Metropolitan, North and South), modality (Public, Private and Subsidized), and curricular branch (Scientific-Humanistic and Technical-Professional).

The evaluation team relied on professional standards for appraising the current PSU process for studying DIF, developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

> **Standard 7.3**
>
> When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability and or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups. (p. 81)

The following subsections contain the results of the evaluation for Objective 1.1.g. Facets 1-2 and DIF demonstration.

Objective 1.1.g. Facet 1. DIF exploratory sources – Document analysis and interviews

GENERAL DESCRIPTION

DEMRE performs differential item functioning (DIF) analysis for the pilot and operational administrations of the PSU tests. DIF analysis is performed for gender (Male and Female) and dependency (Municipal, Subsidized and Private) groups. Gender DIF is explored with males as reference group and females as focal group. Dependency DIF is explored with all possible combinations of categories. There is no policy outlining which of the three sub-groups should be taken as the reference group. Instead, analyses are carried out with each of the group as the reference group.

DEMRE performs DIF analyses with statistical software: SPSS (2010) and the DIFAS program (Penfield, 2007). Both software programs allow the calculation of the DIF by the Mantel-Haenszel method. Additionally, the DIFAS program provides the Breslow-Day statistic. DEMRE relies on the following DIF criteria on the analyses:

- Mantel-Haenszel Chi-Square (MH X^2): The criterion is a critical value of 5.02, which is the value of chi-square with 1 degree of freedom at the alpha = .025 level.
- Mantel–Haenszel Common Log-Odds Ratio (MH LOR): If the value is positive, there may exist DIF favoring the reference group; if the value is negative, there may exist DIF favoring the focal group. Taking into account that mentioned and following the categories used by the Educational Testing Service regarding DIF classification, there are three levels for its interpretation: *irrelevant* (A), *moderate* (B), or *severe* (C).
    - Level A: MHD not significantly different from 0 (based on alpha=.05) or |MHD| < 1.0, where MHD = -2.35 ln(MH LOR).
    - Level C: |MHD| $\geq$ 1.5 and significantly different from 1.
    - Level B: All other cases not encompassed by Level A or Level C. (Allen, Carlson, & Zalanak, 1999)
- (Note: Items showing DIF C are excluded from test assembly efforts and scoring. Moderate DIF items have been used when necessary, but this has only has happened occasionally.)
- Standardized Mantel-Haenszel Log-Odds Ratio (LOR Z): For no DIF to be considered, the value obtained for any item must be in the range of -2.0 to +2.0.
- Breslow-Day by MH X^2 (BD): Generally it matches with the MH X^2 and has the same criteria.

DEMRE performs DIF analyses for each PSU test separately. DIF analyses take place for pilot and operational administrations. DIF analyses have been a part of the DEMRE suite of psychometric quality control tools since 2006. DEMRE documentation of DIF outcomes makes reference to higher DIF rates for the pilot administration than for the operational administration. During test construction, it appears that DEMRE does not put much weight on the results from DIF analyses for pilot items.

The DEMRE research unit analyzes DIF for the pilot and operational administrations. It flows the DIF results to another DEMRE unit that is responsible for maintaining and updating the PSU item bank. The review of items is then a responsibility of the DEMRE test construction unit. This test construction unit reviews items after they have been banked.

EVALUATION

The DEMRE documentation presents how it performs DIF studies, how it processes the data for the analyses and how the results are summarized. The documentation also presents criteria used in classifying items with DIF.

There are a few aspects of DEMRE's approach to DIF analysis that have drawn the attention of the international evaluation team. One aspect is DEMRE's decision to dismiss information concerning the high prevalence of DIF rates for pilot administrations. High rates of pilot DIF manifest significant problems with pilot conditions (e.g., self-selection, differential motivation rates, and high rates of omissions and the representativeness of the pilot sample). These conditions can introduce an element of bias into the total test score and thus affect the matching of ability for comparable groups (Camilli & Shepard, 1994; Dorans & Holland, 1993).

The second aspect of DEMRE's DIF procedures that raises a concern is its decision to emphasize operational DIF results over the pilot DIF results. The international evaluation team considers these decisions problematic because DIF analyses and outcomes are important for assessing fairness proactively during the piloting of items. Addressing DIF through operational results is risky because best practices call for examining the items for bias *before* they are presented to students during operational administration. In an ideal world, quality control processes are set up to detect anomalous items before they become operational. In applied scenarios, retaining items showing DIF requires human expert interpretation of construct irrelevant sources behind the statistical flags (Schmitt, Holland & Dorans, 1993; Zieky, 2006).

A third aspect is the lack of a policy guideline directing selection of reference and focal groups for DIF analyses. The fact that DIF is only calculated for gender and type of dependency is also troublesome. Internationally, the concept of protected classes of test takers has influenced the practice of defining the groups for DIF analyses. Recently the concept was broadened to include variables to better understand factors behind DIF (e.g., curriculum exposure) (Linn, 1993; Schmitt, Holland & Dorans, 1993). DEMRE's analysis should also include such factors as socio-economic status, high school curricular branch and region. Also, we would like to recommend exerting caution on the reliance on multiple comparisons (e.g., Private vs. Municipal, Private vs. Subsidized and Municipal vs. Subsidized) whit not particular hypothesis to check. It is obvious to expect that multiple comparisons could have an effect on the Type I error rate and affect the efficiency of the process. DEMRE's documentation neither provides a rationale nor justifies its use of multiple comparisons. DIF flags are not necessarily indicative of bias; thus, thus DIF analyses cannot be carried out mechanically.

A final aspect of DEMRE's DIF processes that raises concern is its decision to use more than one approach to explore DIF. This approach is reminiscent of the man with two watches who didn't know what the exact time was. When used either individually or conjoined, the approaches aim at detecting DIF relatively to the set of test items, with different levels of statistical error (Type I and Type II) and statistical power (Linn, 1993). DEMRE's documentation does not appear to contain any rationale for using more than one method of DIF analysis neither the relatively superiority of the approaches.

Table 68 shows a summary evaluation for the PSU Selected Operational Item Review and approval (pilot sample and test taker population). The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of identifying a finer grain of information for improvement decisions.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 68: Summary Evaluation of PSU DIF Analyses

Facet 1: DIF exploratory sources – Document analysis and interviews	
1. Describe the process for DIF exploratory analysis	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Is the information found in the technical documentation and is it complete?	C (7.3)

RECOMMENDATIONS

1. We recommend evaluating the significance of pilot DIF results as part of the data review processes and prior to banking the items. Items with DIF C flags should be scrutinized for potential bias by data review panels. Once the items have been analyzed, a record of the decisions reached in data review should be added to the associated item documentation, noting the decision to use or not use the item for operational administration.
2. We recommend expanding DIF analyses to relevant sub-groups that historically have not been part of DEMRE DIF analyses. At a minimum, DIF analyses should be expanded to the following subgroups: region, socio-economic status and curricular branch.
3. We recommend setting a policy for defining reference groups. The process currently followed involves multiple comparisons among categories of the sub-group variable which is not only inefficient but also increases the type I error rate for DIF results.
4. We recommend choosing the Mantel-Haenszel DIF method instead of using multiple DIF methods. The use of Mantel-Haenszel Chi-squared method is well documented and allows for the use of ETS DIF classification rules. If for any reason a backup method is needed, the evaluation team recommends the logistic regression method. The reliance on a single process should be clearly stated in the documentation. The use of multiple methods becomes problematic because different methods have different Type I error rates.
5. Once DEMRE picks a single method for calculating DIF, it should involve content experts to examine those items that have been flagged for DIF. We recommend that DEMRE create criteria for invalidating pilot items with DIF outcomes, such as C flags. The process should differentiate between statistical flagging of DIF and content sources of DIF. The international evaluation team strongly recommends avoiding the use of multiple DIF methods.
6. The evaluation team recommends taking into consideration policy and practical limitations when choosing focal and reference groups for DIF analysis.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Allen, N. L., Carlson, J. E., & Zalanak, C. A. (1999). *The NAEP 1996 technical report.* Washington, DC: National Center for Education Statistics.

Camilli, G. & Shepard, L. (1994). *Methods for identifying biased test items* (Volume 4). Thousand Oaks: SAGE publications.

DEMRE. (2011). *Banco de ítemes.* Santiago: Universidad de Chile.

DEMRE. (2011). *Protocolo de análisis de ítemes, rendición oficial y asignación de puntajes PSU.* Santiago: Universidad de Chile.

Dorans, N. & Holland, P. (1993). DIF detection and description: Mantel-Haenszel and standardization. In (Holland, P. & Wainer, H., Edits). *Differential item functioning* (pp. 35-66). Hillsdale, New Jersey: LEA.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In (Holland, P. & Wainer, H., Edits). *Differential item functioning* (pp. 349-364). Hillsdale, New Jersey: LEA.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Penfield, R. D. (2007). *DIFAS 4.0 user's manual*. Retrieved May 23, 2012, from http://www.education.miami.edu/facultysites/penfield/index.html.

Schmitt, A., Holland, P. & Dorans, N. (1993). Evaluating hypothesis about differential item functioning. In (Holland, P. & Wainer, H., Edits). *Differential item functioning* (pp. 281-315). Hillsdale, New Jersey: LEA.

Zieky, M. (1993). Fairness review in assessment. In (Downing, S. & Haladyna, T., Edits). *Handbook of test development* (pp. 359-376). Mahwah, New Jersey: LEA.

Objective 1.1.g. Facet 2. DIF exploratory sources – Exploration of variables associated with DEMRE's analysis of DIF for 2006-2012

GENERAL DESCRIPTION

Purposes

The purpose of the objective is to assess the prevalence of DIF in the pilot and operational administrations of PSU test items and to explore possible sources of DIF. Analyses were performed with DEMRE DIF results from years 2006 to 2012. Descriptive analyses of DIF magnitude and intensity were summarized by PSU test, year, and type of administration. Inferential analyses were performed with items showing a pair of DIF outcomes (one for the pilot administration and another for the operational administration).

Analyses

Different factors are studied to explore potential sources of differential item functioning (DIF) in the piloting as well as in the final test administration. The contribution of each factor to reduction of variance of Mantel-Haenszel DIF flags is studied with the non-standardized and standardized regression estimates from logistic regression. Additionally the R-squared (R^2) is estimated (square multiple correlation), which expresses the percentage of variance accounted for by the predictor variables.

The analyses are carried out for each PSU test in each year of official application for the Mantel-Haenszel DIF values. Only the items for which statistics have been obtained in the piloting as well as in the official operational application are taken into account. The purpose of the analysis is to summarize the DIF patterns in the pilot and operational test applications.

The analyses are performed using the following statistics:

- CTT difficulty
- CTT biserial correlation
- Percentage of omission
- IRT difficulty
- IRT discrimination

The regression model uses these statistics as independent variables to predict the value of the Mantel-Haenszel DIF statistics.

The analyses are geared toward answering the following questions:

- What is the contribution of characteristics of items, CTT difficulty, CTT biserial correlation, omission percentages, IRT difficulty, and IRT discrimination, on odds of Mantel-Haenszel DIF flags for comparisons between
    - Subsidized vs. Private,
    - Municipal vs. Private,
    - Municipal vs. Subsidized,
    - Female vs. Male?

The analyses were carried out for each test separately and outcomes summarized in tables. Because of space limitations the following name convention was adopted to refer to PSU test identification.

Table 69: Name Convention for PSU Tests

Convention	PSU test name
LC	Language and Communication
M	Mathematics
HSS	History and Social Sciences
S	Science
S-Com	Science – Common Module
S-ElecBio	Science – Elective Module Biology
S-ElecPh	Science – Elective Module Physics
S-ElecCh	Science – Elective Module Chemistry
S-ComBio	Science – Common Module Biology
S-ComPh	Science – Common Module Physics
S-ComCh	Science – Common Module Chemistry

DIF was calculated only when there were data for which 10 or more items flagged; with fewer items, the sample moments matrix is not *positive definite*.

RESULTS

This section is organized into two areas: descriptive and inferential. Descriptive section intends to provide an overlook of the incidence of DIF flags. The second section is geared toward documenting factors that may be associated with DIF flags.

Descriptive Results

In total, 5102 items from all administrations between 2006 and 2012 were analyzed, including items from the pilot and operational administrations. The data were processed with regard to each case of the final administration. The data appear in the following tables. These results were obtained with all of the items.

The percentage of items, per year, marked as FLAG or NON FLAG, for the pilot and for the final administration are shown in Table 70.

Table 70: Percentage of Items Marked as FLAG and NON FLAG in the Pilot and in the Operational Administration

	PILOT		OPERATIONAL	
YEAR	FLAG	NON FLAG	FLAG	NON FLAG
2006	6.52	93.48	3.52	96.48
2007	7.20	92.80	6.09	93.91
2008	11.62	88.38	7.21	92.79
2009	11.19	88.81	7.94	92.06
2010	11.85	88.15	9.44	90.56
2011	10.28	89.72	6.61	93.39
2012	8.28	91.72	9.22	90.78
TOTAL	9.86	90.14	7.23	84.28

In accordance with the procedure described in the DEMRE documents for DIF detection and for the procedure to follow when DIF items are found, it would be expected that the proportion of DIF items in the final administration would decrease to zero with respect to the pilot administration, since in the forming of the final tests, items with DIF in the pilot application or in previous final administrations are not included (or very rarely). However, the table does not indicate a sensible decrease of the proportion of cases, even in the case of 2012, in which the increase in the proportion of cases with FLAG regarding the DIF in the final administration draws attention.

This confirms the use of the pilot DIF information, in the sense that it is not taken into account as a criterion for the selection of the items for the operational administration.

Table 71: Percentage of Items that Repeat Mantel-Haenszel FLAG between Pilot and Operational Administrations

Mantel-Haenszel DIF	FLAG
Subsidized vs. Private	2.61
Municipal vs. Private	3.29
Municipal vs. Subsidized	0.00
Female vs. Male	3.57

The percentage of items that repeat the FLAG condition for DIF between pilot and final shown in Table 71 is much lower than that described in Table 70, indicating that the items change their condition from FLAG to NON FLAG and vice versa between the pilot administration and the final one. On the one hand, this fact could be indicating that the pilot sample, in relation to the DIF verification, is not similar to the final population, and on the other hand, that the values obtained in the pilot administration, in reality do not predict what will happen in the final administration and that it would not be worthwhile to process the data in this sense.

Inferential Results

Table 72 summarizes results, by subpopulation, of regression of Mantel-Haenszel DIF flags on statistical measures of items features. In the table the predictors are listed under the heading predictor and the criterion variable is listed under the heading (Mantel-Haenszel). ** shows statistical significance at .001, and * shows statistical significance at .05. The table shows standardized regression coefficients. The table also shows R2 values. +, ++ or +++ indicate effect sizes are small, medium or great, respectively. If the size of the effect is great it means that the R2 values are significantly different than zero (Faul, Erdfelder, Buchner & Lang, 2009, p. 1155) indicating that there is explanation of the DIF variance through the predictor variables.

Table 72: Regression Weights from DIF Predictors for Each Case Using Mantel-Haenszel

	Mantel-Haenszel DIF			
PREDICTORS	Subsidized vs. Private	Municipal vs. Private	Municipal vs. Subsidized	Female vs. Male
CTT DIFFICULTY	-0.450**	-0.392**	0.019	-0.073**
CTT BISERIAL	0.308**	0.111**	0.046**	0.051**
OMISSION	-0.312**	-0.296**	-0.134**	-0.150**
IRT DIFFICULTY	-0.031**	-0.099**	0.049**	-0.032*
IRT DISCRIMINATION	-0.089**	0.071**	0.045*	-0.056**
R^2	0.403+++	0.268+++	0.025+	0.035+

According to the results, the weights of the predictor variables of the different DIF cases analyzed are significant even though they have differentiating characteristics. In some cases the direction is positive, in others it is negative and in others the weight size is different.

So it is that the CTT difficulty has relatively large and negative weights when the DIF is calculated between Private and another type of school. It is the same for the CTT biserial correlation and omission predictors. On the other hand, the weights of the IRT predictor variables are always relatively small. The direction of the effect of the CTT predictor variables is maintained for all DIF cases analyzed, except for the difficulty in the case of Municipal vs. Subsidized, where it is positive. The same does not happen for the IRT variables whose direction changes in the different cases analyzed. The biserial correlation always has a positive weight, that is, its effect always goes in the same direction of the DIF value.

To conclude, these data allow one to establish that the best DIF predictors, in general, are the CTT variables (difficulty, discrimination and omission) since they are more stable in the sense of the direction and the size of the effect.

Regarding R^2, the table shows greater prediction of Mantel-Haenszel DIF flags for comparisons involving private schools. For example, the set of predictors accounts for by 40% of variance of DIF flags for the subsidized vs. private comparison, and about 27% for the municipal vs. private comparison. Interestingly, the predictors accounted for by less than 4% of variance of DIF flags for the female vs. male comparison and the municipal vs. subsidized comparison.

These results are very interesting since the multiple regression analysis as well as the R^2 (for these grouped data) coincide in identifying the DIF in the cases when whether the school is private or not is analyzed with the other school variables, being this the greatest source of DIF. This is the most stable case among the conditions analyzed. However, it is necessary to take into account that the bias in the data for the case mentioned may take place due to the fact of real differences in the education received by students in both types of schools. In this sense, also, it is important to analyze if the items favor (or disfavor) students from private schools since the pool of item writers may lean toward having more teachers from those types of schools. Upon analyzing in detail the information (documented in the Section Three of the *Psychological Digest*), similar results are found in most cases.

Finally, descriptive analysis of the distribution of DIF flags by ETS category was performed to understand the prevalence of DIF flags on the particular comparisons (see tables 7 to 9).

A common pattern across the tables is the large number of DIF flags categorized as irrelevant, followed by a double-digit number of moderate DIF flags, and the smallest number of flags for severe DIF. This pattern is more pronounced for the Subsidized vs. Private and the Municipal vs. Private comparisons than for the Female vs. Male and Municipal vs. Subsidized comparisons. Interestingly, these last two comparisons showed statistically no significantly R^2. Table 73 through Table 76 show the distributions of Mantel-Haenszel DIF flags by year. Most of the DIF flags are "irrelevant" or "moderate." Of those flags marked "severe," most favor the reference groups. Note that the percentage of items with "severe" DIF is a small percentage of the total number of items examined. This percentage is comparable to what the evaluation team has seen in practice internationally.

Table 73: Distribution of Mantel-Haenszel DIF Flags for Subsidized vs. Private Comparison

YEAR	ETS Level of Significance Subsidized vs. Private	N	Favoring Reference Group*
2006	IRRELEVANT	417	318
2006	MODERATE	12	11
2006	SEVERE	4	4
2007	IRRELEVANT	504	385
2007	MODERATE	30	30
2007	SEVERE	8	8
2008	IRRELEVANT	569	450
2008	MODERATE	53	53
2008	SEVERE	6	6
2009	IRRELEVANT	727	601
2009	MODERATE	71	69
2009	SEVERE	11	11
2010	IRRELEVANT	745	596
2010	MODERATE	109	109
2010	SEVERE	36	36
2011	IRRELEVANT	826	653
2011	MODERATE	62	61
2011	SEVERE	12	12
2012	IRRELEVANT	791	585
2012	MODERATE	102	97
2012	SEVERE	7	7

*Private is the reference group.

Table 74: Distribution of Mantel-Haenszel DIF Flags for Municipal vs. Private Comparison

YEAR	ETS Level of Significance Municipal-Private	N	Favoring Reference Group*
2006	IRRELEVANT	406	294
2006	MODERATE	24	24
2006	SEVERE	3	3
2007	IRRELEVANT	488	355
2007	MODERATE	41	40
2007	SEVERE	13	13
2008	IRRELEVANT	551	435
2008	MODERATE	68	66
2008	SEVERE	9	9
2009	IRRELEVANT	702	570
2009	MODERATE	82	81
2009	SEVERE	25	23
2010	IRRELEVANT	742	560
2010	MODERATE	121	119
2010	SEVERE	27	27
2011	IRRELEVANT	814	610
2011	MODERATE	67	67
2011	SEVERE	19	19
2012	IRRELEVANT	771	578
2012	MODERATE	100	92
2012	SEVERE	29	27

*Private is the reference group.

Table 75: Distribution of Mantel-Haenszel DIF Flags for Municipal-Subsidized

YEAR	ETS Level of Significance Municipal-Subsidized	N	Favoring Reference Group*
2006	IRRELEVANT	433	241
2007	IRRELEVANT	542	270
2008	IRRELEVANT	628	360
2009	IRRELEVANT	809	477
2010	IRRELEVANT	890	472
2011	IRRELEVANT	891	531
2011	MODERATE	9	9
2012	IRRELEVANT	896	566
2012	MODERATE	4	4

*Subsidized is the reference group.

Table 76: Relation between DIF Warning Methods in Female-Male

YEAR	ETS Level of Significance Female-Male	N	Favoring Reference Group*
2006	IRRELEVANT	415	204
2006	MODERATE	14	12
2006	SEVERE	4	4
2007	IRRELEVANT	502	241
2007	MODERATE	25	18
2007	SEVERE	15	15
2008	IRRELEVANT	583	298
2008	MODERATE	36	24
2008	SEVERE	9	8
2009	IRRELEVANT	741	353
2009	MODERATE	42	30
2009	SEVERE	26	26
2010	IRRELEVANT	847	446
2010	MODERATE	33	22
2010	SEVERE	10	4
2011	IRRELEVANT	831	427
2011	MODERATE	47	38
2011	SEVERE	22	22
2012	IRRELEVANT	810	421
2012	MODERATE	80	59
2012	SEVERE	10	10

*Male is the reference group.

Additional tables are available from Section Three (PSU Differential Item Functioning (DIF) Regression Analysis Results — Eight Appendices for Information referenced in Objective 1.1.g) found in the Psychological Digest.

EVALUATION

The DIF analysis is a task that goes beyond the processing of data and the use of standard criteria on those data. The development of procedures that enable the detection of plausible explanations for the presence of DIF should be added to DEMRE repertoire. The evaluation team's analyses of archival PSU DIF data showed a straightforward way to explore for potential sources associated with DIF using logistic regression. This type of analysis could be expanded to accommodate other item attributes, such as the use of words, presence or absence of art and the nature of distractors.

The demonstration provided in this section exemplifies DIF computation for a broader set of demographic or other grouping variables. The follow-up to this demonstration would involve groups of qualified content specialists and teachers in meetings in which items with DIF flags are further analyzed. The outcomes of those meetings are important because they expand the understanding of the factors affecting the quality of items, which can improve item development training and inform item development specifications in the future.

For the pilot, as long the statistical DIF for an item has been analyzed by a bias review committee and these experts have found that the flag was not pointing to any meaningful bias, then the item should be allowed for use if needed to fill content gaps on the test.

For the operational test, if an item is flagged for DIF, it does not necessarily mean the item was biased. A review is needed to determine why the item was flagged. For example, something not found previously (e.g. double keys) and fundamentally wrong (formatting, printing error) could trigger the DIF flag. Omission rates can also muddle the information.

Table 77 shows a summary evaluation for PSU DIF Information Analysis. The purpose of the table is to provide a high level snapshot of expert evaluation of the holistic evaluation of the second facet of the evaluative objective.

In addition to the coding schema shown under column labeled "Rating" in the table, a list of professional standards not met is shown within parentheses.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 77: Summary Evaluation of PSU DIF Information Analysis

Facet 2: DIF EXPLORATORY SOURCES	
1. DIF information analysis	Rating
2. Data from different sources are processed, facilitating the analysis	C (7.3)

RECOMMENDATIONS

1. The PSU program should investigate sources of DIF and use results to fine-tune their item development practices, test construction models, and test scoring process. Analytical approaches can be used to gain understanding on variables that relate to DIF flags.
2. The PSU program should complement the data obtained with DIF detection analysis with the participation of content experts and educators.
3. The PSU program should add other sub-groups currently absent from DIF analyses. At minimum the following sub-groups should be added: region, socio-economic status, and curricular branch. The international evaluation team performed a demonstration of DIF analyses with PSU data from the 2012 admission process. This demonstration could be used as a model for the type of sub-groups to be analyzed in a DIF study.

   This demonstration found, among other results, that:

   - Most of the items showed negligible DIF (A), with very few items showing weak or strong DIF (B or C)
   - PSU Science (common and elective portions) showed larger number of DIF C flags than PSU Mathematics, Language and Social Studies
   - The Gender variable showed the largest number of DIF C flags for a given Common portion of the Science test (six favoring Males and three favoring Females)
   - The variables SES, Region, Curricular Branch, and Modality showed far fewer DIF C flags, with the greatest number of DIF C flags appearing on the Chemistry test for the Scientific-Humanistic versus Technical-Professional curricular branch comparison (three favoring Scientific-Humanistic and four favoring Technical-Professional)

BIBLIOGRAPHY

DEMRE. (2011). *Protocolo de análisis de ítemes, rendición oficial y asignación de puntajes PSU.* Santiago: Universidad de Chile.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behaviour research methods. 41*, 4.

DIF DEMONSTRATION

The international evaluation team performed a demonstration of DIF with PSU data from the 2012 admission process. Differential item functioning (DIF) analysis is a commonly used approach in large-scale assessment geared toward statistically screening out potential item bias. DIF assesses whether an item performs differentially between a focal group and a target (or reference) group of students after matching on their ability. DIF statistics should only be used to identify items showing differential performance after controlling for ability. Subsequent reviews by committees with people familiar with content and students characteristics are required to explore and examine potential sources of performance differences.

DIF analyses were performed with the following sub-groups. Reference groups were defined based on their relevance and sample size. (Note: Sample sizes are available from Table 80 and Table 81.)

Gender:
- Male
- Female
- Reference group: Male

Socio-economic status (SES)[7]:
- Quintile A: Lower
- Quintile B: Below Average
- Quintile C: Average
- Quintile D: Above Average
- Quintile E: Upper
- Reference group: Quintile A & B[8]

Region:
- Central
- North
- South
- Reference group: Central

Modality:
- Municipal
- Subsidized
- Private
- Reference group: Municipal

Curricular branch:
- Scientific-Humanistic
- Technical-Professional
- Reference group: Scientific-Humanistic

[7] The five SES levels were computed with the SES procedure that accounts for family income and parental education.

[8] The first two quintiles were combined to make the reference group for our analysis. The evaluation team chose the first two quintiles because in its opinion they are a reasonable point of reference for comparing the other quintiles of the SES distributions. Different results could be reached if each quintile had been dealt with separately.

DIF analyses were conducted with the frequency table approach and Mantel-Haenszel statistic using Pearson proprietary engine (e-PRS). The Educational Testing Service (ETS) DIF classification schema was followed to flag items for any of the following categories:

- Category A (Negligible DIF)
- Category B (Weak DIF)
- Category C (Large DIF)

Table 78 summarizes the number of PSU items for the 2012 admission process by PSU Mathematics, Language, and Social Studies and DIF category (A, B, C). Note that for the PSU tests most of the items fall within category A, followed by a small number of items with DIF category B, and a very small number of items with DIF category C. Across the three content areas, PSU social studies showed the largest number of DIF flags for categories B and C, followed by PSU Language and PSU Mathematics.

Interestingly the modality variable captured all the items with large DIF (category C) for the comparison between Municipal vs. Private high schools. Further analyses showed differences favor private high schools (Mathematics and language) and municipal schools (social studies).

Table 79 summarizes the number of PSU items for the 2012 admission process by common and elective portions of PSU Science assessment and DIF category (A, B, C). Most of the items showed negligible DIF (e.g., DIF category A), followed by small number of DIF B and C flags. Interestingly, PSU Science (common and elective portions) showed larger number of DIF C flags than PSU Mathematics, Language and Social Studies.

The gender variable showed the largest number of DIF C flags for a given Science test. All of these nine flags observed were found for the common portion of the Science test. From the nine flagged items, six favored Males and three Females. The curricular branch variable showed the largest number of DIF C flags for the entire Science test. The flags were concentrated for Chemistry (seven flags), followed by Physics and the Science common portion (three flags each), and Biology (one flag). The DIF C flags were not found systematically favoring either of the two curricular branches. For example, for Chemistry out of the seven DIF C flags, three favor the Scientific-Humanistic curricular branch and four favor the Technical-Professional curricular branch. For Physics, out of the three DIF flags, two favors Scientific-Humanistic and one favor Technical-Professional. The modality variable showed the smallest number of DIF C flags among variables with DIF C flags. One item was flagged for Physics. The DIF C flag for that item favored the private high school group.

Table 78: Mantel-Haenszel DIF Result for PSU Mathematics, Language and History and Social Studies from the Year 2012 Administration

Sub-groups	Mathematics			Language			Social Studies		
	A	B	C	A	B	C	A	B	C
Gender:									
- **Male** vs. Fem	70	4		74	4		64	9	2
SES:									
-QA& B vs. QC	74			78			75		
-QA& B vs. QD	74			77	1		75		
-QA& B vs. QE	74			75	3		75		
Curricular branch:									
-**Scientific** vs. technical	68	6		77	1		74	1	
Modality:									
-**Municipal** vs. subsidize	74			78			75		
-**Municipal** vs. private	73		1	72	5	1	69	4	2
Region:									
-**Metropolitan** vs. South	74			78			74		1
-**Metropolitan** vs. North	74			78			74	1	

The **boldface** categories are the reference groups.

Table 79: Mantel-Haenszel DIF Result for PSU Science, Common and Elective, from the Year 2012 Administration

Sub-groups	Science (Common)			Biology (Elective)			Chemistry (Elective)			Physics (Elective)		
	A	B	C	A	B	C	A	B	C	A	B	C
Gender:												
- **Male** vs. Fem	32	12	9	25	1		23	1	2	26		
SES:												
-QA& B vs. QC	53			26			26			26		
-QA& B vs. QD	53			26			26			26		
-QA& B vs. QE	53			26			25	1		25	1	
Curricular branch:												
-**Scientific** vs. technical	47	3	3	23	2	1	16	3	7	20	3	3
Modality:												
-**Municipal** vs. subsidize	53			26			25	1		26		
-**Municipal** vs. private	51	2		25	1		25	1		24	1	1
Region:												
-**Metropolitan** vs. South	53			26			26			26		
-**Metropolitan** vs. North	53			26			25		1	26		

The **boldface** categories are the reference groups.

Table 80: Distribution of n-Counts for DIF Analyses for Language, Mathematics and History and Social Sciences

	Language		Mathematics		History and Social Sciences	
Comparison	N	%	N	%	N	%
Gender:						
- **Male**	53891	46.65	50257	46.73	30922	47.35
- Female	61639	53.35	57286	53.27	34390	52.65
SES*:						
- **QA&QB**	40914	38.05	40698	37.95	26684	41.02
-QC	26889	25.00	26843	25.03	16345	25.13
-QD	17541	16.31	17527	16.34	9844	15.13
-QE	22191	20.64	22186	20.69	12179	18.72
Curricular branch:						
-**Scientific-Humanistic**	81416	71.14	75731	71.09	43038	66.50
-Technical-Professional	33025	28.86	30794	28.91	21678	33.50
Modality:						
- **Municipal**	41764	36.49	38626	36.26	24508	37.87
-Subsidized	60323	52.71	56108	52.67	33523	51.80
-Private	12354	10.80	11791	11.07	6685	10.33
Region:						
-**Central**	59658	51.68	55754	51.88	34437	52.75
-North	12963	11.23	11938	11.11	7038	10.78
-South	42823	37.09	39773	37.01	23808	36.47
Test Length	80		75		75	
Number of Invalidated Items	2		1		0	

The **boldface** categories are the reference groups.
***The "unevenness" of the quintile percentages is due to the coarseness of the SES scale.

Table 81: Distribution of n-Counts for DIF Analyses for Science (Common and Electives)

	Elective Biology		Elective Chemistry		Elective Physics		Common Science	
Comparison	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Gender:								
- **Male**	12622	35.58	6543	46.64	9946	77.12	31107	46.46
- Female	22857	64.42	7487	53.36	2951	22.88	35846	53.54
SES*:								
- **QA&QB**	14135	39.94	3771	26.92	3520	27.34	21524	34.49
-QC	8937	25.25	3262	23.29	2887	22.43	15104	24.20
-QD	5914	16.71	2821	20.14	2430	18.88	11177	17.91
-QE	6405	18.10	4154	29.65	4037	31.36	14601	23.40
Curricular branch:								
-**Scientific-Humanistic**	26754	76.20	12100	86.88	9891	77.34	52350	78.92
-Technical-Professional	8356	23.80	1828	13.12	2898	22.66	13981	21.08
Modality:								
- **Municipal**	13274	37.81	4136	29.70	3870	30.26	23005	34.68
-Subsidized	18884	53.79	7668	55.05	6751	52.79	35744	53.89
-Private	2952	8.41	2124	15.25	2168	16.95	7582	11.43
Region:								
-**Central**	16284	45.93	6598	47.07	6796	52.76	31719	47.42
-North	3677	10.37	2051	14.63	1660	12.89	7977	11.92
-South	15495	43.70	5367	38.29	4426	34.36	27199	40.66
Test Length	26		26		26		54	
Number of Invalidated Items	0		0		0		1	

The **boldface** categories are the reference groups.

*****The "unevenness" of the quintile percentages is due to the coarseness of the SES scale.

Objective 1.1.h. Analysis of procedures for the calculation of standardized scores, score transformation in relation to the original distributions

Procedures for scaling and norming tests are well documented in literature, and there are diverse methods available to practitioners (Kolen & Brennan, 2004). The PSU relies on several scales along the process to standardized test scores. The process begins with a raw scale that has an origin of zero and a maximum score equal to the number of items in the test. All numbers in the scale are integer numbers. Another scale is the raw score corrected for guessing scale. Raw scores are corrected by taking one quarter of incorrect scores out of raw scores. Item omissions are considered as not reached and thus do not count against applicants' raw scores. The scale for the corrected-for-guessing score is a primary scale ranging from a floating negative integer to a maximum number of items in the test. The scale rounds corrected raw scores utilizing a rule of equal or greater than 0.75. For example, whereas a corrected score of 68.45 in Mathematics rounds to 68; a score of 56.75 rounds to 57 points. A third scale is the normalized score scale, which is derived by the inverse function of relative cumulative frequency distribution of corrected raw scores from a given test administration. This scale ranges from -5 to +5 with a precision of four decimal points. The fourth scale is the PSU scale with a mean of 500 points and a standard deviation of 110 points. This scale is a linear transformation of the normalized scale, and it has been truncated to have a minimum score of 150 and a maximum score of 850 points. An example of a PSU scale score is 612.35 points.

Additionally to derivation of scale scores, the PSU also involves a smoothing of the PSU scale score distribution. The purpose of the smoothing is to soften distribution bumps in the upper portion of the PSU scale up to 1% of the distribution. The process is performed manually, and it involves a linearization of the scores utilizing simple ratios (proportions).

In addition to the scaling of PSU test scores, a scaling of *Notas de la Enseñanza Media* (NEM) took place for the 2003 admission process (DEMRE interview). The outcome of that process was a set of normative tables that connect NEM (on the raw scale) to the NEM standardized score. The normative tables were developed for Scientific-Humanistic (morning and afternoon) and Technical-Professional curricular branches. The NEM scale has a mean of 500 and a standard deviation of 100 points. Such a scale was inherited from properties of the PAA scale.

The evaluation team performed the interviews with relevant stakeholders from DEMRE on March 23 of 2012. The interview process took a total of 2 hours from 14:00 to 16:00. The purpose of the interview was to gain deeper understanding on:

- Types of scales, standardized scores and their computational process
- Process to transform PSU scores in relation to the original distribution

All the interviews were performed within DEMRE offices following an agreed-upon schedule for the visit. The interviews covered the two facets and relevant elements agreed upon with the TC during the goal clarification meeting in Santiago, Chile, in January 2012.

The following DEMRE staff participated in the interviews:

- Head of research unit and his team

- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees.

The international evaluation team relied on professional standards for the appraising the merit and worth of the PSU process for developing scale scores. A framework for evaluating PSU approaches for scale scores and norming has been developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA NCME, 1999).

Standard 2.2

> The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretations. (p. 31)

Standard 2.14

> Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (p. 35)

Standard 4.1

> Test documents should provide with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations. (p. 54)

Standard 4.2

> The construction of scales used for reporting scores should be described clearly in test documents. (p. 54)

Standard 4.5

> Norms, if used, should refer to clearly described populations. These populations should include individual or groups to whom test users will ordinarily wish to compare their own examinees. (p. 55)

Standard 4.6

> Reports of norming studies should include precise specification of the population that was sampled, sampling procedures, and participation rates, any weighting of the sample, the dates of testing, and descriptive statistics. The information provided should be sufficient to enable users to judge appropriateness of norms for interpreting the

scores of local examines. Technical documentation should indicate the precision of the norms themselves. (p. 55)

Standard 4.8

When norms are used to characterize examinee groups, the statistics used to summarize each group's performance and the norms which those statistics are referred should be clearly defined and should support the intended use or interpretation. (p. 56)

The following subsections contain the results of the evaluation for Objective 1.1.h., Facets 1-2.

Objective 1.1.h. Facet 1. Types of scales, standardized scores and calculation procedures

GENERAL DESCRIPTION

By policy emanated from *Consejo de Rectores de Universidades Chilenas* (CRUCH), the reporting scale for the PSU test scores, which is still in use, was developed for the first time in 2004 and finalized. Until 2003 a reporting scale known as the PAA scale was used to report scale scores on the admission test. The PAA scale was a standardized scale with an average of 500 points and a standard deviation of 100 points. In the first PSU test (2004) some of the PSU tests were reported utilizing the PAA scale and others utilizing the new PSU scale. The PSU scale is a standardized scale with a mean of 500 points and a standard deviation of 110 points. Policy from the CRUCH formalized the use of the PSU scale for all the PSU tests for the admission process of year 2005 and beyond.

Scaling is a process to assign numbers (or other ordered characteristics) to intended attributes such as test performance and high school performance. Once a scale has been developed, year-to-year equating of test scores allows maintaining the scale. In Chile, PSU test performance is reported with the PSU scale but there are no efforts put in place to maintain the scale across admission years.

In Chile, the process to develop PSU scale scores is comprised by multiple steps ranging from computation to PSU raw score to computation of applicants' postulation scores. As mandated by the CRUCH, the purpose of the PSU tests is to select students for pursuing studies in careers offered by the CRUCH and the eight private Chilean universities that have subscribed to it. That is, the PSU test scores are used to rank order applicants seeking admission into university-level studies.

The following definitions describe the scores and their relationship to the computation of applicants' postulation scores.

- Raw Score: number of correct answers to multiple-choice items scored 1/0.

- Corrected for Guessing Score: from the total correct answers, subtract one fourth of the wrong answers. The corrected score is rounded up when the non-integer portion of the score fraction part is at least 0.75. The correction for guessing is intended to increase the motivation of qualified applicants to attempt at answering the PSU items. In our evaluation of this practice, we found the use of the correction for guessing questionable (see the Evaluation section).

- Normalized Score. This kind of score is computed from inverse of the relative cumulative frequency distribution observed for the corrected for guessing raw score. The reliance of a normalized score is a common practice for norm-referenced contexts seeking to rank order applicants. It enables presenting percentiles and percentile ranks to allow interpretation of PSU raw scores.

- PSU scale score. This type of score is calculated from a linear transformation of the normalized score. The transformation forces the distribution to take a mean of 500 and a standard deviation of 110. The PSU scale truncates lower and upper portions to achieve a minimum score of 150 and a maximum score of

850, respectively. Linearly transforming the normalized scores is a policy consideration. Because the linear transformation does not change rank ordering of applicants, percentile and percentile rank information can be associated to the PSU scale. For example, by setting the population of applicants mean equal to 500 scale score points, the PSU test score reported to an applicant would indicate whether that examinee is above or below the population mean.

For a typical university admissions organization, the process of scale development ends with the production of raw-to-scale score conversion tables for reporting test scores to test takers and university admissions officers, for example. In Chile, DEMRE goes beyond the production of the raw-to-scale score tables to support the admission process by computing and providing the additional scores below.

Weighted PSU scale scores by the university or the career. Each one of the universities or careers assigns weight to the PSU tests counting for the selection process. While the selection of the weights is a decision made within each career, the computation of the weighted PSU scale score is a core responsibility of DEMRE.

Postulation score. This score is computed by combining the weighted PSU scale scores and the standardized high school average score (NEM score). While the selection of the weights to combine PSU and NEM responds to an admission policy mandated by the CRUCH, the computation of the postulation score is a core responsibility of DEMRE. From the interviews, the evaluation team learned about NEM since DEMRE has not documented the process to develop the NEM norms.

EVALUATION

The international evaluation team recognizes efforts made to provide PSU standardized test scores and their corresponding computational procedures. Collectively, the steps followed for computing PSU scale scores resemble the canonical structure of processes for computing scale scores in norm-reference contexts (Petersen, Kolen, & Hoover, 1989). The steps respond to admission policy mandated by the CRUCH and DEMRE's role as a data processing entity for the postulation scores. As mandated by the CRUCH, PSU test scores are used to select students for university studies. The PSU test scores are used to rank-order applicants seeking admission into college level studies.

The documentation of PSU processes to develop scale scores (partial completion of professional standard 4.1) provides a good starting point but this documentation needs more elaboration and inclusion of supporting evidence for: (1) correcting for guessing, (2) rationale for choosing mean and standard deviation of PSU scale scores, (3) decisions for truncating the PSU scale scores, and (4) maintenance of PSU scale. In addition, the evaluators recognized serious problems in the use of formula scoring adopted for PSU. Most of the professional standards listed for evaluating this facet were not fulfilled either in the PSU technical documentation presented or in the follow-up inquiries that were made in the form of interviews (See professional standards 2.2, 2.14, 4.2, 4.5, 4.6, 4.8). The set of deficiencies can be remedied in future years.

The international evaluation team regards the use of the correction for guessing as inadequate because it challenges the validity of PSU scores and PSU field test administration results. Correction for guessing, also known as formula scoring, is a

traditional approach to addressing differential guessing among examinees in classical test theory framework. Formula scoring is based on a model of applicants' response to an item. Three possible situations are covered by the model: (1) test taker knows the correct answer, (2) test taker omits the item, test takers choose randomly an option answer (Rowley and Traub, 1977). Two versions of formula scoring are available from literature: (1) rewarding omissions and (2) punishing guessing. Although these two approaches render different scores, applicants' rank order remains unchanged. That is, formula scoring approaches yield scores that are perfectly correlated. Nevertheless, Traub, Hambleton, and Singh (1969) warned practitioners about the potential differential effects that these approaches could have on test takers' behaviors.

Formula scoring has been criticized for its unrealistic assumptions (Frary, Cross & Lawry, 1973, Diamond & Evans, 1973) and for its weak representation of behaviors in the real world (Gullikese, 1950; Lord and Novick, 1968). Test takers' behaviors are not necessarily consistent with the assumptions of formula scoring. For example, formula scoring does not take into consideration, as it should, the partial knowledge of test takers. Additionally, formula scoring does not boost test score reliability and test score prediction validity as was initially expected. The formula for guessing also creates unnecessary complexities on calculations of quality measures, such as test score reliability. Moreover, the use of formula scoring has a negative effect on public opinion, which might judge the process as punitive for students. Finally, the use of formula scoring may end up changing the balance of PSU test blueprint due to the fact that test takers, in their effort to avoid being penalized with 0.25 for each wrong response, could omit items; which, when treated as not presented, could affect content representation of PSU. For all of the reasons stated above, the international evaluation team does not support the use of formula scoring.

These problems with formula scoring have been recognized elsewhere. For example, in the U.S. the most influential university admissions programs have either, as in the case of the ACT, historically avoided the use of correction scores or, as in the case of the SAT, have begun migration efforts out of its use.

There is a caveat that applies to the PSU Science test that requires of additional discussion. The Science test has two modules: (1) a common and (2) an elective. Applicants take the common module and must take one of the three elective modules (Biology, Physics or Chemistry). In all the tests forming part of the PSU battery, questions are multiple-choice, with five alternative answers, only one of which is correct. The total Science raw scores are obtained adding the result of the common module to the optional one. This latter one has an adjusted scale with a linking procedure, taking as reference one of the three optional modules. The base module is not necessarily maintained throughout time, which has become a concern for the evaluating team. For more information on the evaluation that has been carried out on the Science test linking, please see Objective 1.2 of this report.

The precision of the scale in which results are reported is not estimated. The test reliability and standard error of measurement is estimated from raw scores utilizing classical test theory. Precision of scale scores both typical and conditional is not part of DEMRE processes. This is a serious limitation considering decisions are made on the scale score metric. Conditional standard errors of scale scores involved in the university admissions decisions should be calculated and communicated to PSU audiences. For more information on the evaluation of the PSU reliability and conditional standard error please see Objective 1.1.i of this report.

The greatest concern on the development of the PSU scale is the lack of a mechanism put forward for maintaining the PSU scale across years. As mentioned before, in Chile, PSU test performance is reported with the PSU scale but no equating takes place to maintain the scale across admission years. This is a serious issue that must be attended to ensure Chile university admissions test is fair to test takers and ensure valid comparisons of test scores across test administrations. For more information on the evaluation of equating of the PSU, please see Objective 1.3 of this report.

Finally, from the interviews, the evaluation team learned about high level features of the process to standardize the high school grade point average (NEM). NEM is computed by averaging final grades attained in high school studies. The grading system ranges from a minimum score of one point to a maximum score of seven points. To the date of the interview, DEMRE was not able to deepen our understanding by providing more information on the structure of the NEM and the process to develop NEM norms because of the lack of documentation existing on those subjects. Due to the role NEM plays in the computation of the postulation scores, it is troublesome not knowing psychometric properties of the NEM scores. Of special significance is the comparability of meaning of NEM test scores across multiple subpopulations. NEM scores are based on grading practices that may or may not be comparable across schools (public, private, and subsidized) and curricular branches (Scientific-Humanistic and Technical-Professional), for example.

Table 82 shows a summary evaluation for PSU scales and standardized scores. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing more detailed information for improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 82: Summary Evaluation of Types of Scales, Standardized Scores and Calculation Procedures

Facet 1: Types of scales, standardized scores and calculation procedures	
1. Describe the types of scales of the PSU, standardized scores and their computer processes	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
a. Describe the main and collateral PSU. With respect to the main scale, let us understand the scale over which decisions are made. • With respect to collateral scale, let us understand the scales supplementing additional pieces of information. • What considerations were given to the predicted use of the PSU test scores (criteria reference vs. guideline reference; selection vs. placement vs. account rendition) and predicted population of test takers when choosing main and collateral scales?	A (4.1,4.5, 4.6, 4.8)
b. Which is the classification of the PSU scores (for example, raw, adjusted regarding guessing, percentile and smoothed)? • What considerations were provided to the predicted use of the PSU test scores (criteria reference vs. guideline reference; selection vs. placement vs. account rendition) and predicted test taker population?	A (4.1, 4.5, 4.6, 4.8)
c. What was the recommendation for the use of PSU standardized scores? • What type of score admission decisions are based upon? • What approaches are followed for estimating the precision of that score?	A (4.1, 4.5, 4.6, 4.8)
d. Describe the rationality behind the processes for PSU standardized score derivation (for example, 500 average and 110 standard deviation. Cutoff of the distribution tails). • What type of information supported the selection decision with respect to the origins of the PSU scale and dispersion unit? • What process is followed for maintaining the PSU scale score throughout the years? I hear that equating of PSU test scores is not used. Why? Could you please clarify what type of process is followed to compensate with respect to differences in difficulty of the forms between previous and current administrations?	D*
e. Do the predicted PSU scores allow comparison with other standardized scores used commonly?	D*

• Do you find other instances to inform about other types of standardized test scores apart from those already informed?	
f. Describe the rationality behind the use of correction with regard to guessing. I know we have already discussed the issue of correction through guessing in other parts of the interview. Here, my intention is to know more about the interaction between correction through guessing and the PSU standardized test scores.	D*

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

1. In Chile, the contribution made by the correction for guessing process to the improvement of PSU test score reliability, PSU predictive validity of scores, and PSU public opinion is not documented. Internationally, the use of the correction for guessing faces challenges in these areas as presented in the summative evaluation of this facet. Additionally, when omissions are considered as not reached, the corrected by guessing scores vary for students with the same number correct score with different omission rates. The correction for guessing may lead the students to use a strategy for approaching the test which does not have to do with their knowledge, thereby reducing the accuracy of the prediction of university performance. Finally, in light of international standing of the correction for guessing on university admissions programs, the international evaluation team recommends considering abandoning the practice of correcting for guessing for future administrations.
2. We recommend considering item response theory as an alternative approach to deal with applicant's guessing behavior. Current PSU scoring approaches use correction for guessing processes that have been found with severe limitations in the literature. As mentioned above, the process adds layers of complexity to multiple aspects of the processes such as when calibrating pilot testing responses, computing item statistics and test score statistics. The item response theory framework brings build in features to account for the amount of guessing (i.e., pseudo-guessing) present in test taker's item responses. We also recommend that in preparation to transition out from the correction for guessing context, if decided, DEMRE prepares and submits a transition plan to an external expert group of reviewers. The plan among other aspects should involve risk analyses and feasibility analyses and a time frame to introduce the necessary changes on critical processes of the university admissions testing program such as PSU test construction, PSU item banking, PSU pilot, PSU scale maintenance, PSU validity and reliability, PSU score reports. The international evaluation team also recommends a series of retrospective studies to evaluate any potential effects on historical trends of PSU test scores and item banking field test statistics, for example, of the decisions made in the past due to use of formula scoring.
3. Because of the inclusion of the standardized high school grade point average (NEM) into the postulation score, the evaluation team recommends the evaluation of their normative data. Conversion tables for NEM were first used in the 2003 admission process. Because NEM is one of the two elements defining the postulation score for most of careers with the CRUCH and the eight affiliated private universities, we recommend deepen the information available

on the structure of NEM and the process to compute it. Along these lines we recommend studying the validity of the inferences drawn from the set of normative data that is almost ten years old and that is based on a national curriculum that has been almost replaced by the current national curriculum. Along these lines of research and documentation, the international evaluation team recommends studying the validity of standardized NEM scores. Of special significance is comparability of meaning of NEM test scores. NEM scores are based on grading practices of unknown generalizability across type of schools (public, private, and subsidized) and curricular branches (Scientific-Humanistic and Technical-Professional). The international evaluation team proposes studying the possibility of replacing NEM with standardized measures of high school academic performance such as scores from nationally administered tests. It is of crucial to properly balance the PSU test frameworks and their reference to Chile's national curriculum to avoid over-emphasizing measures of high school academic performance while sacrificing measures of general scholastic aptitudes.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2005). *Procedimientos para el cálculo de los puntajes estándar.* Santiago: Universidad de Chile.

DEMRE. (2011). *Protocolo de análisis de ítemes, rendición oficial y asignación de puntajes PSU.* Santiago: Universidad de Chile.

DEMRE. (25 de Agosto de 2011). *Fórmulas y consideraciones: Aprende a calcular los puntajes*. PSU en el Mercurio. Documento No. 8. http://www.demre.cl/text/publicaciones2012/agosto/publicacion11(25082011).pdf

Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research, 43,* 181-191.

Frary, R., Cross, L., & Lowry, S. (1977). Random guessing, correction for guessing, and reliability of multiple-choice test scores. *The Journal of Experimental Education, 46*(1), pp. 11-15.

Gilliksen, H. (1950). *Theory of mental tests.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Lord, F., & Novic, M. (1968). *Statistical theories of mental test scores.* Reading MA: Addison Wesley Publishing Company.

Pearson Evaluation Team. (2012). PSU evaluation interviews.

Petersen, N., Kolen, M., & Hoover, H. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3 rd ed., pp. 221-262). New York: Macmillan.

Traub, R., Hambleton, R., & Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement, 29*, 847-861.

Objective 1.1.h. Facet 2. Score transformation in relation to the original distributions

The evaluation team performed the interview with relevant stakeholders from DEMRE on March 23, 2012. The interview process took a total of 2 hours from 14:00 to 16:00. The purpose of the interview was to gain deeper understanding on:

- Types of scales, standardized scores and their computational process
- Process to transform PSU scores in relation to the original distribution

The interview was carried out following an agreed-upon schedule. The interviews covered the two facets and relevant elements agreed to by the TC during the goal clarification meeting in Santiago, Chile, in January 2012.

The following DEMRE staff participated in the interviews:

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

As part of this objective, Pearson provided hands-on computation of PSU corrected raw scores, PSU scale scores, and PSU smoothed scale scores (See companion *Psychometric Digest* for the results of these analyses). The analyses seek to compute PSU standardized scores and score transformation in relation to the original distribution of corrected raw scores. For these analyzes DEMRE processes were followed when documentation available to the evaluation made such processing feasible. The analyses were performed for the most recent operational administration of the PSU (i.e., 2012 admissions process). Results were reported separately for each of the PSU test in an appendix of the evaluation technical report. When sample sizes were large enough to guarantee stable results, PSU standardized scores were summarized with descriptive statistics by each of the following set of variables: socio-economic status (SES), region (Metropolitan, North and South), modality (Public, Private and Subsidized), and curricular branch (Scientific-Humanistic and Technical-Professional).

The international evaluation team relied on professional standards for appraising the merit and worth of the PSU process to transform scales. A framework for evaluating PSU approaches for scale transformation is developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 2.2

> The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretations. (p. 31)

Standard 2.14

> Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (p. 35)

Standard 4.1

> Test documents should provide with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations. (p. 54)

Standard 4.2

> The construction of scales used for reporting scores should be described clearly in test documents. (p. 54)

Standard 4.5

> Norms, if used, should refer to clearly described populations. These populations should include individual or groups to whom test users will ordinarily wish to compare their own examinees. (p. 55)

Standard 4.6

> Reports of norming studies should include precise specification of the population that was sampled, sampling procedures, and participation rates, any weighting of the sample, the dates of testing, and descriptive statistics. The information provided should be sufficient to enable users to judge appropriateness of norms for interpreting the scores of local examines. Technical documentation should indicate the precision of the norms themselves. (p. 55)

Standard 4.8

> When norms are used to characterize examinee groups, the statistics used to summarize each group's performance and the norms which those statistics are referred should be clearly defined and should support the intended use or interpretation. (p. 56)

GENERAL DESCRIPTION

In Chile, admission decisions are made at the career level within a university with a postulation score, which combines selected PSU scores and *Notas de la Enseñanza* Media (NEM). From the interviews, the evaluation team learned high level features of the process to standardize the high school grade point average (NEM). NEM is computed by averaging final grades attained in high school studies. The grading system ranges from a minimum score of one point to a maximum score of seven points.

The postulation score provides a single index of applicants' academic achievement over PSU content areas deemed relevant to careers and the record of academic performance in high school measured with the high school grade point average (NEM). As described in the previous facet, each university career is autonomous in deciding on the set of weights for combining PSU scores and NEM scores, while retaining compliance to the mandated admission policy from *Consejo de Rectores de Universidades Chilenas* (CRUCH). In Chile, predictive validity studies have been conducted by scrutinizing the relationship between PSU test scores, NEM and first-year college grade point averages (*Comité Técnico Asesor*, 2010). These studies have noted the Pearson product-moment correlation coefficients between PSU test scores and NEM and first-year college grade point averages. Findings have shown that PSU Mathematics and PSU Science tests result with larger predictive validity coefficient than NEM. In contrast, PSU Language and History tests resulted with the lowest predictive validity (*Comité Técnico Asesor*, 2010). The evaluation team found a pattern that is similar to the one described above (See Objective 2.4 in this report). The correlations between NEM and PSU scores are documented in Appendix M. Weighted Correlations of NEM and PSU Subtest Raw Score.

PSU test developers also report standalone PSU scores and NEM scores. Whereas PSU test scores are reported with both scale scores and percentile ranks, NEM scores are reported with scale scores. The description of these processes follows:

> The process to compute PSU test scale scores begins with a raw scale that has an origin of zero and a maximum score equal to the number of items in the test. All numbers in the scale are integer numbers. The process continues by referencing the previous scale to another scale that accounts for correction for guessing. Raw scores are corrected by taking one quarter of incorrect scores. Item omissions are considered as not reached and thus do not count against applicants' scores. The corrected-for-guessing scale ranges from a floating negative integer to a maximum value equal to the total number of items in the test. The scale rounds corrected raw scores utilizing a rule of equal or greater than 0.75. For example, whereas a corrected score of 68.45 in Mathematics rounds to 68; a score of 56.75 rounds to 57 points. The correction for guessing of a raw score is not interpreted since its primary function is to provide input to the score transformation process.
>
> To develop norm reference interpretations of PSU scores, corrected-for-guessing scores are normalized first. The normalization process is performed in such a way that the shape of the original distribution is transformed while retaining the original rank-ordering of applicants. Then, normalized scores are linearly transformed to the PSU scale with a mean of 500 points and a standard deviation of 110 points. This scale is a linear transformation of the normalized scale, and it truncates below 150 and above 850 scale score points. An example of a PSU scale score is 612.35 points. The normative interpretation built into the PSU scale allows for a type of interpretation such as those made with percentile ranks. For example, an applicant scoring 500 scale score points on the PSU Mathematics test is scoring better than approximately half of the population of applicants.
>
> In addition to the derivation of PSU scale scores; the PSU also involves a smoothing of the PSU scale score distribution. The purpose of the smoothing is to soften bumps in the upper portion of the PSU scale up to 1% of the distribution. The process is performed manually, and there is no documentation available in DEMRE documentation. During the interviews, DEMRE staff mentioned, after being

prompted by interviewers, that simple ratios (proportions) are used to smooth the distribution of scores at the upper 1% region.

From the interviews, the evaluation team learned about high level features of the process to standardize the high school grade point average (NEM). The average high school score is computed by averaging concentration raw grades in high school studies. The high school grading system ranges from a minimum score of one point to a maximum score of seven points; with four points defining passing on the scale. To the date of the interview, DEMRE was not able to deepen our understanding by providing more information on the structure of the NEM and the process to develop NEM norms because of the lack of documentation existing on those subjects. Due to the role NEM plays in the computation of the postulation scores, we delved into DEMRE Internet portal (DEMRE, 2011a) to gain more information on NEM processes and interpretations. The following paragraphs bring summary of the information with our comments.

The process to compute NEM scale scores relies on the use of NEM norms developed for the 2003 admission process. The NEM normative tables are available for three groups, separately:

- Group A: Scientific-Humanistic (Morning). It comprises high school graduates from Scientific-Humanistic schools attending morning classes, and graduates from the navy.

- Group B: Scientific-Humanistic (Evening). It comprises high school graduates from Scientific-Humanistic evening schools, graduates with partial studies abroad, and graduates with validation tests.

- Group C: Technical-Professional. This group comprises graduates from commercial, industrial, technical, agricultural and maritime areas.

College applicants' high school grade point averages –a number ranging from 4 to 7, allowing for a decimal point (e.g., 5.0)– are read into the relevant set of norms to find their corresponding NEM scale scores. The average high school score was computed by taking the average of high school subject score (DEMRE, 2011). The NEM set of norms were developed in 2003 through linear transformations of the standardized high school average scores, which was performed separately for each of the three groups (see Table 83).

The NEM scale scores are reported onto a scale with a mean of 500 points and a standard deviation of 100 points. Such characteristics of the scale (e.g., multiplicative and additive constants) were inherited from the PAA scale features.

The transformation process of standardized applicants' high school average scores to NEM standard scores relies on a linear function with a multiplicative coefficient of 100 points and an additive coefficient of 500 points. For example, a college postulant from group *A* showing an average high school score of 5.0 points attains a NEM scaled score of 414 points. Along the same lines, a postulant from group *B* with an average high school score of 5.0 points attains a NEM scaled score of 417 points. (Note: to the date of the interview, DEMRE was not able to deepen our understanding by providing more information on the structure of the NEM and the process to develop NEM norms (e.g., n-counts, means and standard deviations of the normative groups) because, as they expressed, there is no documentation on those subjects.)

Since the scores within a group (e.g., group A, B or C) are referenced to the mean and standard deviation attained by the normative group (e.g., Group A in 2003), NEM scaled scores can be used to make within-groups comparison. Between-group comparisons, on the other hand, are not feasible to make by looking at the difference on NEM scaled scores between the groups. Unless the groups showed comparable distributions of average high school raw scores, the derived NEM scaled scores cannot be comparable.

Table 83: Transformation Tables between Average High School and NEM Scaled Score

SCORE	GROUP A SCIENTIFIC-HUMANISTIC MORNING	GROUP B SCIENTIFIC-HUMANISTIC EVENING	GROUP C TECHNICAL-PROFESSIONAL
4.0	208	218	213
4.1	229	238	233
4.2	249	258	254
4.3	270	279	274
4.4	290	299	295
4.5	311	319	315
4.6	332	339	335
4.7	352	359	356
4.8	373	380	376
4.9	393	400	397
5.0	414	420	417
5.1	435	440	437
5.2	455	460	458
5.3	476	481	478
5.4	496	501	499
5.5	517	521	519
5.6	538	541	539
5.7	558	561	560
5.8	579	582	580
5.9	599	602	601
6.0	620	622	621
6.1	641	642	641
6.2	661	662	662
6.3	682	683	682
6.4	702	703	703
6.5	723	723	723
6.6	744	743	743
6.7	764	763	764
6.8	785	784	784
6.9	805	804	805
7.0	826	824	825

(DEMRE, 2011a)

Internationally, university admissions processes incorporate multiple sources of information on the academic performance of applicants over and above the scores from entrance examinations. High school ranking and performance on national standardized tests (or statewide graduation tests) are examples of measures tapping on past academic performance. In Sweden, for example, grades from national centrally administered tests in core subject areas (Swedish, English, Mathematics, and Chemistry) have played a role on university admission processes (Stage, 2003). The use of the centrally administered tests, instead of relying on classroom grades, is one potential avenue to discount for heterogeneity of classroom grading practices.

EVALUATION

The international evaluation team recognizes DEMRE efforts to compute scale scores for postulation scores, PSU test scores, and NEM scores. Nevertheless, the documentation of the process for developing scale scores for those reports is partial and incomplete, thus only partially accomplishing the recommendations for professional standard 4.1. The preliminary documentation of the processes provides a good starting point, but several scales and their processes are not documented (e.g., postulation score scales, smoothing PSU test scores, NEM norms). There also needs to be more elaboration on ways to define measurement precision for standalone PSU scale scores and their composite with NEM. The above deficiencies can be improved in the near future.

The evaluation team also recognizes serious problems in the documentation and technical adequacy of scales, which has caused the team to disapprove the associated processes for scale development. Most of the professional standards listed for evaluating this facet were not fulfilled (professional standards 2.2, 2.14, 4.2, 4.5, 4.6, 4.8). These deficiencies can be improved in the near future.

Neither the psychometric properties nor a proposed method of interpretation of the postulation score have not been documented. While admission criteria stipulates use of PSU test scores and NEM scores (along a set of weights and policy considerations) as building blocks of the postulation score, little is known about the psychometric characteristics of the composite postulation score such as its scale mean and unit of dispersion. While a normative meaning is attached to the individual PSU test scores, evidence for the meaning attached to the postulation score has not been defined.

Another main problem we detected is the lack of information on measurement precision of the postulation scores. When ranking postulation scores, numerical differences between postulation scores are of no consideration despite their potential lack of practical significance. For example, a score of 672.15 points ranks above a score of 672.13 points; nevertheless, the two scores show differences at the second decimal place. Without measures of score precision available to understand differences in the scores, it is plausible that differences of the size of a few decimal points could be interpreted and used to make decisions on who gets admitted and who does not. In summary, a lack of conditional standard error of measurement for postulation scores is a major drawback that needs to be addressed in the near future. More discussion on needs to compute measures of accuracy and precision is included in the section covering Objective 1.1.i (reliability and conditional standard error of measurement).

The evaluators found some level of documentation of the descriptions of the processes to derive PSU scale scores for individual PSU tests. The reviewed documentation and information from interviews helped in understanding the process to develop the scale

and the process to assign meaning to scale points. For example, while documentation summarized PSU scale characteristics, interviews helped the evaluations obtain fine-grain information on the PSU scale. DEMRE documentation needs to be expanded to cover description of measurement precision (e.g., conditional standard error of measurement) for individual PSU tests, which to the present date has been neither computed nor reported to users. It is also recommended that summaries of the limitations of derived scale scores be provided, when applicable. For example, we found the process for manually smoothing the upper tails of the distributions (1%) problematic for the following reasons. The processes followed by DEMRE depend on human judgment and lack of quality control checks. In addition, DEMRE has not coined evidence on the effects of the smoothing on bias reduction.

The evaluators found little to no documentation of the processes for deriving norms for *Notas de la Enseñanza Media* (NEM). During interviews with DEMRE staff, international evaluators uncovered the existence and use of NEM norms. The interviews with relevant stakeholders made evident that NEM norms have been used since 2003. From the interview it was also learned that there are three sets of NEM norms: (1) Scientific-Humanistic (morning attendance), (2) Scientific-Humanistic (afternoon attendance), and (3) Technical-Professional, respectively. Throughout the interviews, DEMRE staff expressed their lack of knowledge of the conditions under which the norming was performed and commented on the lack of availability of technical reports on the norming studies. At the time of the interviews, DEMRE was not able to deepen our understanding by providing more information on the structure of the NEM and the process to develop NEM norms because of the lack of documentation. Due to the role that NEM plays as part of the computation of the postulation scores, the lack of information on NEM norms and the psychometric properties of the NEM scores results are troublesome. The lack of evidence supporting comparability of classroom grading practices is another concern. NEM scores are based on grading practices that may or may not be comparable across schools (Public, Private and Subsidized) and curricular branches (Scientific-Humanistic and Technical-Professional), for example.

Table 64 shows a summary evaluation for PSU scales and standardized scores in relation to normalization and smoothing of scores. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing aspects for fine grain improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 84: Summary Evaluation of Score Transformation in Relation to the Original Distributions

Facet 2: Score transformation in relation to the original distributions	
1. Describe the methods used for deriving scales and transforming scores and their rationality.	
2. Follow-up questions (in case they are necessary and if applicable)	Rating
b. Does the transformation take place for postulation scores and PSU test scores?	A (4.1, 2.2, 2.14)
c. Describe the process to smoothen the original PSU test score distributions. • What part of the policy and research informed the decision to smoothen the original score distribution of the PSU tests? • About what types of needs is the smoothing dealing with? How in its process is the smoothing dealing with the conditional standard errors? • What considerations were provided to the predicted use of the PSU test scores (criteria reference vs. guideline reference; selection vs. placement vs. accountability) and regarding the predicted test taker population on smoothing the original distribution of the PSU scores? • What other alternatives to incorporate meaning to the scale scores were inspected before deciding upon the election of the smoothing approach?	A (4.1, 4.2, 2.2, 2.14, 4.5, 4.8)
d. Describe rationality behind the smoothing process (for example, objectives distribution, number and types of smoothing parameters, manual / automatic, penalized / not penalized).	A (4.1, 4.2, 2.2, 2.14, 4.15, 4.8)
e. Which are the advantages and disadvantages of its use (for example, conditional measuring error)?	A (4.1, 4.2, 2.2, 2.14)
f. Describe the implementation of the smoothing process.	A (4.1, 4.2, 2.2, 2.14)
g. What software is used for carrying out the smoothing? How much documentation is there on both the process and the	A (4.1, 4.2, 2.2, 2.14)

smoothing software?	
h. Describe the process followed to produce NEM scale scores. • Characteristics of populations • Update process across time • Technical report for NEM norming	A (4.1, 4.2, 4.5, 4.8, 2.2, 2.14)

RECOMMENDATIONS

The international PSU evaluation team thinks the foundations of scales and reported standardized scores are set to a barely minimal level, which, when conjoined with the decision to correct scores by guessing (a process that models unrealistically students test taking behavior), leads the evaluation team to reject the facet. Based on this judgment, the evaluation team proposes instituting the following processes to improve the program.

1. We recommend completely revising the postulation score system using the perspective of composite scores. Procedures for computing composite scores and their associated scale scores are well documented in the literature, and there are diverse methods available to practitioners (Feldt & Brennan, 1989; Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). These efforts should be instituted to provide precise information at those regions of the postulation score scale where important decisions are made (e.g., admission decisions), while acknowledging for the differential composite weights used across universities and their careers.
2. We recommend introducing the measurement precision of reported standardized scores for individual tests into the PSU test score and postulation score system. The PSU program should develop guidelines on intended uses and interpretation of PSU scale scores and standardized scores with an emphasis on delineating the limitations of the use and interpretation of derived scores. The efforts should keep in perspective the differential composite weights used across universities and their careers.
3. We recommend providing technical documentation of the norming of NEM scores with an emphasis on descriptions of the intended PSU test populations, sampling procedures, participation rates, weighting approaches (if used), testing dates, and descriptive information of background variables.
4. We recommend maintaining a research agenda to study year-to-year stability of primary and secondary PSU scales.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2005). *Procedimientos para el cálculo de los puntajes estándar*. Santiago: Universidad de Chile.

DEMRE. (2011a). *Documentos técnicos*. Retrieved from http://www.demre.cl/escala_p2011.htm

DEMRE. (2011b). *El promedio de notas de enseñanza media es de gran importancia: No descuides tus notas*. PSU en el Mercurio. Documento No. 11. Retrieved from http://www.demre.cl/text/publicaciones2012/septiembre/publicacion14(15092011).pdf

Feldt, L., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd edition, pp. 105-146). New York: American Council on Education and Macmillan.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Norwell MA: Kluwer Academic Press.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Stage C. (2003). *Entrance to higher education in Sweden*. Paper presented at School of Education, University of London.

Objective 1.1.i. Reliability (CTT) and precision (IRT), including the information function, of the different instruments forming part of the PSU test battery - Standard error analysis of conditional measurement for the different score distributions sections, placing special emphasis on the cut off scores for social benefits

Information on reliability and measurement error is essential to the proper evaluation of test scores and their use. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) defines reliability as "consistency of measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25). The standards defines measurement error as "the hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true or universe score for the procedure" (p. 25).

Certain types of test score use require of scores with less measurement error than other score uses. When test scores are used to make academic decisions over the test score range, a test that is capable to discriminate equally well over the test score scale is seen as a more tenable option than a test showing selective discrimination on a particular region of the test score scale. On the other hand, when test scores are used with particular cut scores points to make decisions (e.g., pass/fail), the amount of measurement error from the test should be somewhat minimized at those scores.

It is important to consider the goals of the university admissions program when evaluating reliability approaches in the context of PSU test scores. In Chile, PSU test scores are used as part of the university admissions criteria to rank order college applicants for selection purposes. The design of the PSU tests reflects two characteristics: reliance on a domain defined by Chile's national high school curriculum and use of the test scores to rank-ordering college applicants for selecting those above of a defined expectation. In Chile, the development of the PSU test (e.g., test construction, test scoring, and test scaling) is directed by classical test theory principles.

The central reliability question for the use of the PSU tests scores is: what is the likely margin of error present in the reported test scores? In the professional literature, there are several measures of test score accuracy and precision within classical test theory (CTT) and item response theory (IRT) frameworks. The following set shows three kinds of indices that are typically generated for educational tests: (1) reliability, (2) conditional standard errors of measurement, and (3) classification accuracy. Reliability measures gears at estimating accuracy and precision of test scores. Conditional standard error of measurement estimates amounts of measurement error at particular test scale points. Classification accuracy seeks to document the likelihood of classification of applicants (into admitted/no admitted categories) if they had taken a parallel version of the assessment under the same administration conditions and about the same time.

The evaluation team developed and performed the interviews with relevant stakeholders from DEMRE on March 23 of 2012. The interview process took a total of 2 hours from 16:15 to 18:15. The purpose of the interview was to gain a deeper understanding on the:

- Process for estimating PSU test score reliability from CTT and IRT frameworks
- Process to compute conditional standard errors for PSU scale scores

All the interviews were performed within DEMRE offices following an agreed-upon schedule for the visit. The interview covered the two facets and relevant elements agreed with the TC.

The following DEMRE staff participated in the interviews:

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees.

This objective also provides a demonstration of the computation of conditional standard error of measurement (CSEM) of PSU test scores under item response theory framework (IRT). DEMRE does not compute nor report the above measure of accuracy as part of their processing of admission test data. DEMRE provides reliability and standard error of measurement from classical test theory. The demonstration used data from the 2012 PSU admissions process. Processes and results of the demonstration are presented at the final section of the summary for Objective 1.1.i. The complete set of results of the demonstration is available from the psychometric digest.

The international evaluation team has relied on professional standards during its appraisal of the merit and worth of the PSU process for estimating test score reliability and conditional standard error of measurement. A framework for evaluating PSU approaches for reliability and standard error of measurement has been developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 2.1.

> For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information function should be reported. (p. 31)

Standard 2.2.

> The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation. (p. 31)

Standard 2.4.

> Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported. (p. 32)

Standard 2.11.

> If there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard errors of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended. (p. 34)

Standard 2.14.

> Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. When cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score (p. 35)

Standard 2.15.

> When a test or combination of measures is used to make categorical decisions, estimates should be provided of percentage of examinees who would be classified in the same way on the two applications of the procedure, using the same form or alternate forms of the instruments. (p. 35)

GENERAL DESCRIPTION

Technical reports on PSU reliability (DEMRE, 2010, 2011) acknowledge a primary use of scores from the PSU battery: specifically, to order candidates for university admissions from highest to lowest so that those who have a higher probability of academic success in college will be selected. Implicit in this declaration of the intended use is a specific understanding of a PSU score and the binary nature of decisions made with the PSU score. In Chile, the university admission criterion is defined as a weighted composite of PSU scores and high school grade point average (NEM). The PSU test scores are reported on a scale with an average of 500 and a standard deviation of 110 points. The high school grade point average (NEM) is reported on a scale with an average of 500 and a standard deviation of 100 points.

Universities affiliated with the CRUCH assign weights to each PSU test for their career-based degree programs in anticipation of DEMRE's computation and reporting of applicants' weighted university admission scores. Applicants participating in a selection process must choose the elective test(s) to be taken according to the requirements set and disseminated by each university for admission into the career(s) they are interested in. The CRUCH has provided guidelines on the lower and upper bound of weights for combining elements of the admission criteria; and universities and their career centers are solely responsible for choosing the specific weights within the allowed range of permissible weights that DEMRE will ultimately use to compute the applicants' weighted university admission score.

DEMRE uses coefficient alpha to estimate the reliability of PSU scores.

> Para el propósito de las PSU® el coeficiente de confiabilidad que se determina es el de equivalencia, pues interesa conocer el grado de precisión de los puntajes obtenidos de su aplicación, y el procedimiento estadístico aplicado corresponde al coeficiente Alfa de Cronbach (1951) el que refleja el grado en el que covarían los ítemes que constituyen el test. (DEMRE, 2010, p. 11)

The purpose of the PSU reliability analyses in the report is to examine reliability coefficients for the full set and selected subsets of applicants to establish whether the test has equivalent levels of efficiency for relevant intended subpopulations. The use of admissions tests requires that precision of measurement be analyzed in subpopulations to see whether attained score reliability has a positive impact on test results and decisions. A finding of different levels of reliability would suggest that there are problems with the test battery requiring further attention. If a test is undergoing preliminary development, modest reliabilities such as 0.70 are acceptable. For general surveys of the population and comparisons of group means, as in some kinds of analytical studies, reliabilities of 0.80 are appropriate. This level of reliability is not nearly high enough for making decisions about individuals. When selection standards are quite rigorous, as they are in university admissions, a reliability of 0.90 should be regarded as a bare minimum, with 0.95 considered an appropriate standard.

The reliability report includes 180 tables presenting reliability coefficients for national totals and selected subgroups of applicants (by gender, region, school administrative dependency, and admissions year) for components of the PSU battery. For the major tests, these coefficients are high (0.95) and surprisingly consistent, with little variation among groups. Lower coefficients are found for some of the smaller subsets of items within tests. It should be noted that standards to appraise the magnitude of estimated reliability coefficients depend on how a measure is going to be used. On this basis, results presented in the present study would appear to satisfy the general requirements for university admissions. However, we cannot be entirely certain that this is so. For one thing, reliability is, by definition, true-score variance as a proportion of total (error plus true-score) variance. One could say that it refers to the population as much as it does to the measure. As such, it refers the overall quality of measurement in a given situation.

EVALUATION

It is a tenet of sound test development and use to document reliability, standard error of measurement, conditional standard error of measurement at each score and their combination into a single composite score. When evaluating reliability estimates of PSU test scores, it is important to consider the use and interpretation of PSU test scores. Certain score uses require greater confidence in the accuracy of the test than other score uses. For example, in Chile, important decisions are made with PSU test scores such as granting university admissions and granting scholarships. If these decisions are to be properly executed, they must attend to the characteristics of university admissions and scholarship granting process. Although there is a unique set of cut scores orienting decisions along the two venues, most decisions tend to be made on the test scale score region above the center of the PSU scale (500 points). When cut scores are used, the amount of information the test produces should be somewhat maximized at those scores, particularly when high-stakes decisions are being made.

The reliance on internal consistency measures (e.g. coefficient alpha) and classical test theory standard error of measurement provides a partial coverage of what is expected for a high stake assessment test from international standards (AERA/APA/NCME, Standard 2.2).

The process to estimate reliability focuses on reporting estimates for individual PSU tests; nonetheless, admission decisions are made with composite scores weighting PSU individual test scores and the high school grade point average. The report does not provide a rationale for skipping the reporting of reliability of the composite university admissions score, which is

ultimately the piece of information upon which university admission decisions are made. In addition to the lack of an estimate of reliability of the university admission composite, there are no pieces of information on the measurement error for the composite score or on the confidence bands around the reported associated percentiles.

The process to estimate reliability involves typical formulation of coefficient alpha for number-right multiple-choice test. In Chile, PSU test scores are produced with a formula scoring that (1) takes out from a correct multiple-choice response one fourth of a point from every multiple-choice wrong response and (2) leaves unaffected the number right score when applicants omit the item. Although DEMRE relies on the above "corrected by guessing" observed score when computing PSU test scores, the computed PSU reliability coefficients are based on PSU raw scores. Interestingly, raw scores do not account for guessing correction. Expressions for reliability coefficients accommodating correction for guessing scoring formulas can be involved in the study of PSU reliability estimates (Frary et al., 1977).

The scope of the PSU reliability estimates that we found is limited when providing a rationale for relying on just number correct coefficient alpha and ignoring estimates of classification consistency/accuracy estimates. When continuous scores are interpreted with respect to one or more cut scores, the coefficient alpha and the standard error of measurement may produce information that may be unrelated to the following question: "How consistent is pass/fail classification?" Since the primary use of the PSU test scores is to screen between applicants to university careers reaching the admission score and applicants not reaching the admission score, accuracy of classifications is an important piece of information that ought to be included as part of the report, if the audience for that report is to understand the degree of decision consistency achieved. Such pass/fail decisions are better informed with classification consistency and classification accuracy approaches which are standard psychometric practices involving a single administration (Livingston, 1972; Subkoviak, 1976; Huynh, 1976).

National and international standards recommend using the standard error of measurement as a gauging principle to compare groups instead of simply comparisons of reliability estimates. It is well known that reliability estimates are group dependent whereas measurement error is not. The reliability report provides information about the amount of measurement error for typical applicants (standard error of measurement). However, as far as the measures themselves are concerned, it would be better to consider measurement precision. Technically, this is the inverse of the error variance of individual measures. In classical test theory (CTT), standard error of measurement assumes that error is the same over the entire assessment scale. It is more realistic to assume that standard errors are smaller at proficiency levels where large numbers of items are concentrated. This implies that precision is concentrated at the middle of a proficiency distribution and lower in the tails of that distribution where relatively few items are found. When standard errors are plotted in relation to proficiency, this produces a U-shaped curve.

For university admissions examinations, it is important that the center of this U-shaped curve is positioned over the range in the proficiency distribution where admissions decisions are likely to be made. One can think of this position as a certain range of percentiles within the examinee population. Adding historical information on the percentages of examinees admitted to careers to the measures of PSU test score precision, universities and their careers could include additional information to orient their selection decisions.

Because of the importance of the postulation score on the admission process, the fact that NEM scale score and PSU scale score differ in their dispersion creates concern. With NEM

reported on a scale with narrower dispersion (i.e., a standard deviation of 100), variance of NEM in the postulation score would be small and this scaling decision would enter into the variance of the postulation score and its reliability. Postulation scores are weighted composites involving PSU test scores and NEM and adopted set of weights. The variance of postulation scores is a function of (1) the squared weighted variance of PSU test scores and NEM scores and (2) the weighted covariance PSU test scores and NEM scores. The international evaluation team recommends addressing the scaling of NEM for future administrations and when computing current norms for such variable. A sensible scaling decision would be to reset NEM scale to use that which is similar to the PSU scale as part of NEM norming activities.

Table 85 and Table 86 show summary evaluations for PSU reliability and precision, respectively. The purpose of the tables is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing aspects for fine grain improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 85: Summary Evaluation of PSU Score Reliability

Facet 1: PSU Score Reliability	
1. Describe processes for estimating PSU test score reliability from CTT and IRT perspectives.	
2. Follow up questions (if needed and if applicable)	Rating
a. What model(s) are used to estimate reliability (e.g., CTT / IRT)? • I am hearing IRT is not used, so what is used? On what type of scores (e.g., raw, corrected for guessing, scale score, weighted score for admission, etc.) reliability is computed? What consideration(s) were given to PSU primary and accessory scales when selecting the approaches to estimate reliability? • What consideration(s) were given to the use of correction for guessing when selecting the approaches? • What is the unit of analysis (e.g., individuals, careers, schools)	A (2.1, 2.2, 2.4, 2.14, 2.15)
b. Describe the processes for estimating measures of precision (e.g. reliability) of PSU test scores. • What considerations were given to the type of score interpretation (norm vs. criterion referenced interpretation) when selecting these approaches?	A (2.1, 2.2, 2.4, 2.14, 2.15)
c. What considerations were given to the declared use of PSU scores (e.g., selection) when choosing reliability processes?	A (2.1, 2.2, 2.4, 2.14, 2.15)
d. What standards are used for evaluating the magnitude of reliability coefficients?	A (2.2, 2.4, 2.14)
e. Is precision estimates disaggregated by sub-group (e.g., gender, socioeconomic status, region, type of establishment, curricular branch, career)?	A (2.2, 2.4, 2.11)
f. How measures of score accuracy disaggregated by sub-groups are interpreted and reported to PSU audiences (e.g., applicants, university admissions officers, high schools)?	A (2.2, 2.4, 2.11)
g. How reliability information is built in the reporting of the PSU scores? • For what types of PSU scores are built in?	A (2.1, 2.2, 2.4)
h. How frequent the processes and criteria for estimating PSU test score reliability are evaluated by an external advisory committee? What is the composition of the committee? How committee's recommendations are implemented and follow-up?	D*

***A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

Table 86: Summary Evaluation of PSU Conditional Standard Error of Measurement

Facet 2: PSU Conditional Standard Error of Measurement	
1. Describe processes for estimating PSU conditional standard errors around critical cut points.	
2. Follow up questions (if needed and applicable)	Rating
a. What model(s) are used to estimate conditional standard errors (e.g., CTT / IRT)? I am hearing IRT is not used, so what model is used? On what type of scores (e.g., raw,	A (2.1, 2.2, 2.4, 2.14,

corrected for guessing, scale score, weighted score for admission, etc.) conditional standard error is computed? • What consideration(s) were given to PSU primary and accessory scales when selecting the approaches to estimate conditional standard error of measurement? What consideration(s) were given to the use of correction for guessing when selecting the approaches?	2.15)
b. Describe the regions of the scale where major decisions (e.g., selection, granting social benefits) take place. • To what degree conditional standard error of measurement informs major decisions? • How were the regions chosen and how much research informed the process? • To what extend the process was reviewed externally? If yes, to what extend recommendations were attended?	A (2.1, 2.2, 2.4, 2.14, 2.15)
c. Description of the estimation processes of conditional standard errors, and justification for their use. • What considerations were given to the type of score interpretation (norm vs. criterion referenced interpretation) when selecting these approaches? • To what extend the process was reviewed externally? If yes, to what extend recommendations were attended?	A (2.1, 2.2, 2.4, 2.14, 2.15)
d. Rationalization of the estimation process regarding the use of scores (e.g., selection, granting social benefits, school accountability).	A (2.1, 2.2, 2.4, 2.14, 2.15)
e. To what extent conditional standard error of measurement is disaggregated by sub-groups of interest (e.g. gender, socioeconomic status, and region, type of establishment, curricular branch, and career)? • Is this information reported to intended users of PSU test scores?	A (2.1, 2.2, 2.4, 2.14, 2.15, 2.11)
f. For what sort of PSU score (raw, corrected for guessing, scale score, weighted score for admission) reporting, conditional standard error of measurement is reported?	A (2.1, 2.2, 2.4, 2.14, 2.15)
g. To what extent interpretation of meaning of conditional standard error is provided to the intended PSU audiences? • To what degree interpretation of processes and outcomes are tailored to audience's need?	A (2.1, 2.2, 2.4, 2.14, 2.15)
h. Do you mind stating the criteria involved in judging the magnitude of the conditional standard error of measurement? • How frequent the processes and criteria for estimating PSU conditional standard error of measurement is evaluated by an external advisory committee? • What is the composition of the committee? How committee's recommendations are implemented and follow-up?	A (2.1, 2.2, 2.4, 2.14, 2.15, 2.11)

RECOMMENDATIONS

The international PSU evaluation team is inclined to think that the reliability coefficients presented in this study are too blunt an instrument to detect whether there are problems with the test battery that require further attention. A research agenda for PSU may include further exploration of ways to report classification consistency/accuracy for composite scores. A worthwhile start would be an initial exploration of approaches to estimate classification consistency/accuracy for individual PSU test scores, followed by research studies to undertake the task to estimate classification consistency/accuracy for composite PSU test scores. It is also important to explore avenues on how to communicate consistency/accuracy of pass/fail decisions and corresponding admission policy to relevant stakeholders of the admission process. The research agenda may take parallel strides to cover secondary uses of PSU test scores such as assignment of scholarships. An agenda focusing on the following areas of improvement would be prudent.

1. The PSU reliability report is limited when providing justification of the approaches to estimate reliability that have been used. In the analyses reported, the coefficient alpha was implemented to estimate reliability. Coefficient alpha shows specific sources of measurement error relevant for some type of decisions. Specifically, coefficient alpha shows measurement error due to sampling error associated to items.
2. Discussion of plausible systematic sources of errors on PSU scores is also absent from the PSU reliability report. The report is lacking discussion of effects and treatment to accommodate correction for guessing and omission. The effects of these two conditions deserve more study. Similar challenges were noted for the estimation of the standard error of measurement, which relied on the standard deviation of uncorrected raw test scores.
3. The PSU reliability report is limited when providing information on conditional standard errors and on the rationale for establishing the acceptable size of such errors for the intended primary (placement) and other (scholarship) uses of PSU test scores among the expected populations of test takers. Of particular interest is the reliance on reliability estimates to explain the consistency of test scores across groups of applicants. The sections dealing with these group comparisons may be better served by descriptions of standard error of measurement that are less group-dependent.
4. The PSU reliability report is limited when providing descriptions of the amount of measurement error at critical regions of the PSU scale utilized to make high-stakes decisions (e.g., accepted/rejected admission and granted/not granted scholarship). Measures of decision consistency/accuracy are important pieces of information currently absent from the estimate of PSU score reliability and precision. Additionally, the existing processes do not explain the precision of PSU scores for primary and other decisions (e.g., university admission and granting scholarships, respectively). The *Standards for Educational and Psychological Testing* (AERA, APA, NCME) advise reporting precision of the scores on the scale from which decisions are made.
5. We recommend addressing the scaling of NEM for future administrations and when computing current norms for such variable. A sensible scaling decision would be to reset NEM scale to use that which is similar to the PSU scale as part of NEM norming activities.

DEMONSTRATION OF CSEM

The purpose of this section is to provide a demonstration of the computation of conditional standard error of measurement (CSEM) of PSU test scores under item response theory framework (IRT). DEMRE does neither compute nor report the above measure of accuracy as part of their data processing activities. Instead, DEMRE computes and documents classical test theory measures of reliability and standard error of measurement. The demonstration we have provided relies on analyses performed on random samples of applications taken from the population of applicants from the 2012 PSU admissions process. The evaluation team chose sample sizes that were large enough (approximately 10,000 cases) to provide stable estimates for the item parameters and related IRT measures. The demonstration made use of the three parameter logistic formulation (3PL) to model correct/incorrect responses to the PSU multiple-choice items with difficulty, discrimination and pseudo-guessing parameters.

This section includes the following test-level and item-level IRT graphs for each PSU test:

- The latent ability distribution
- The test characteristic curve (TCC)
- The test information function (TIF)
- The conditional standard error of measurement (CSEM)

Each of these graphs is first introduced with a realistic companion example. The PSU Mathematics test was chosen to illustrate the outcomes from the demonstration followed by illustration of outcomes for the other PSU test before introducing descriptions for the other PSU tests. A more comprehensive set of demonstration results can be found from the Psychometric Digest.

A look-up table was developed to correlate ability measures (expressed in normal deviates) and their linearly transformed scale scores (mean = 500 and SD = 110).

Table 87: IRT Ability Scale and Scale Score (Mean=500 and SD=110)

IRT Ability	Scale Score
-3.0	170
-2.5	225
-2.0	280
-1.5	335
-1.0	390
-0.5	445
0	500
0.5	555
1.0	610
1.5	665
2.0	720
2.5	775
3.0	830

Disclaimer: This table shows a linear transformation of IRT ability which is not intended to report PSU scale scores based on IRT ability estimates.

Latent ability distribution

The latent ability distribution depicts the density of ability estimates for a random sample of students (n=8909) considered for the demonstration. It shows an estimate of where the target population of examinees is located in relation to ability or proficiency estimates, where proficiency is measured as normal deviates (mean=zero and stdev=1). Normal deviates, like percentiles, are measures of relative position in a distribution of scores. The position and extent of the latent ability distribution is used to determine whether tests, for example, (1) are appropriate for the target population, (2) are off-target low and too easy or (3) are off-target high and too difficult.

Figure 11 shows the latent population distribution for PSU Mathematics test under the 3-PL formulation adopted for the demonstration. The horizontal axis depicts proficiency on PSU Mathematics on the theta scale and the vertical axis shows proportion of test takers. The density of the latent ability distribution shows a bell shape and deviates ranging from about -3.0 to +2.75 with mean latent ability at zero and standard deviation of 1.00.

Judging by the distribution of applicants' latent ability, the ability spans within a reasonable range for an admission test with a domain based on high school curriculum. For example, it becomes apparent the larger number of applicants with ability scores below one standard deviate than the number of applicants with ability scores above two standard deviates. A score of at least 610 points (one standard deviation above the mean of 500 points) is attained by a reasonable fraction of applicants.

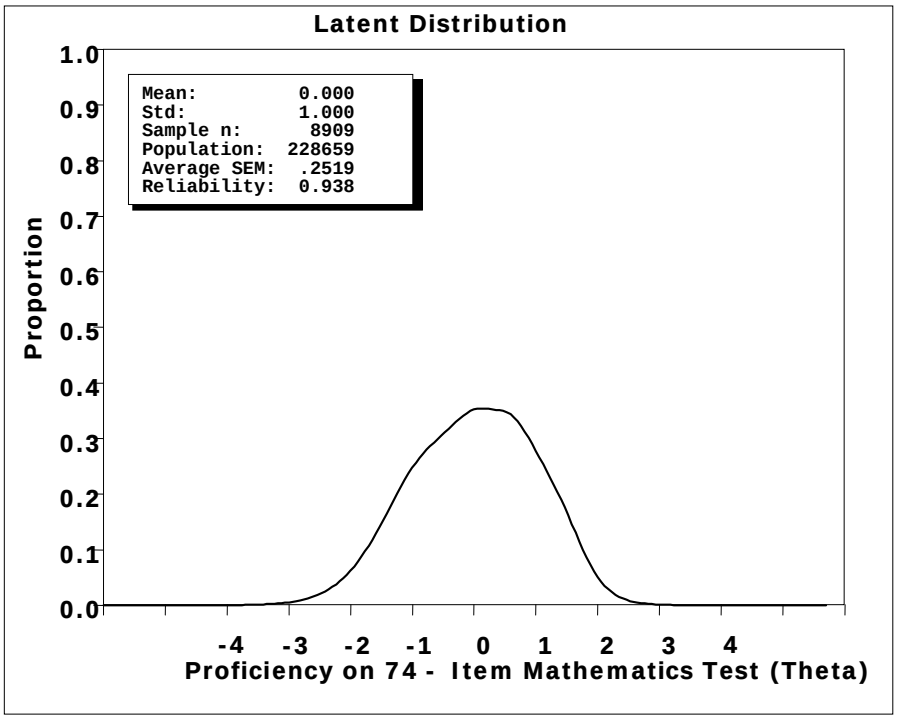


Figure 11: Latent Ability Distribution for 74-Item Mathematics Test

Figure 12 through Figure 16 show latent ability distributions for PSU Language and Communication, History and Social Sciences, Biology, Physics, and Chemistry tests, respectively. Characteristics of these distributions are comparable to those characteristics highlighted for the Mathematics test and collectively show distributions of ability expected to be found on a test informing norm-referenced interpretations.

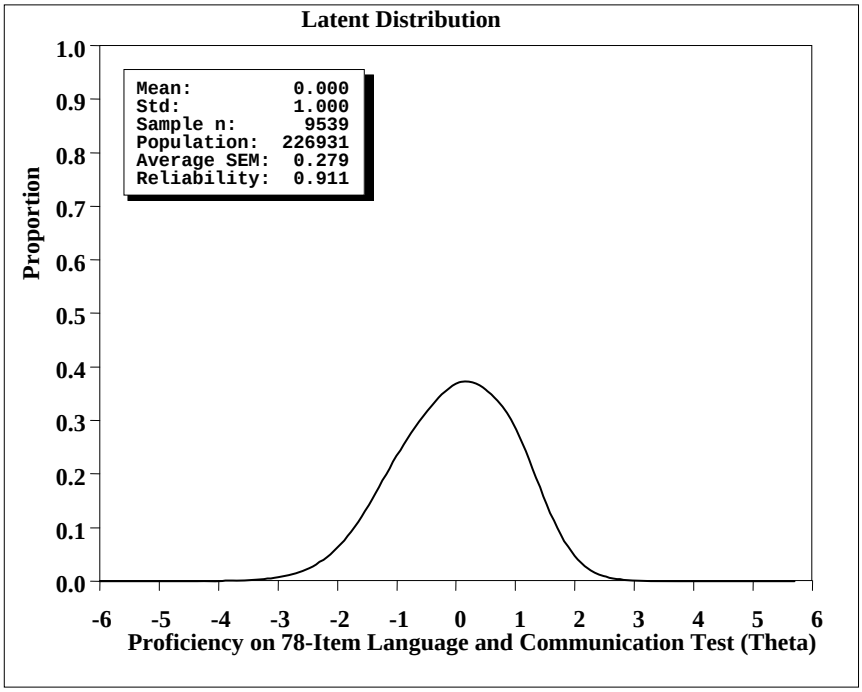


Figure 12: Latent Ability Distribution for 78-Item Language and Communication Test

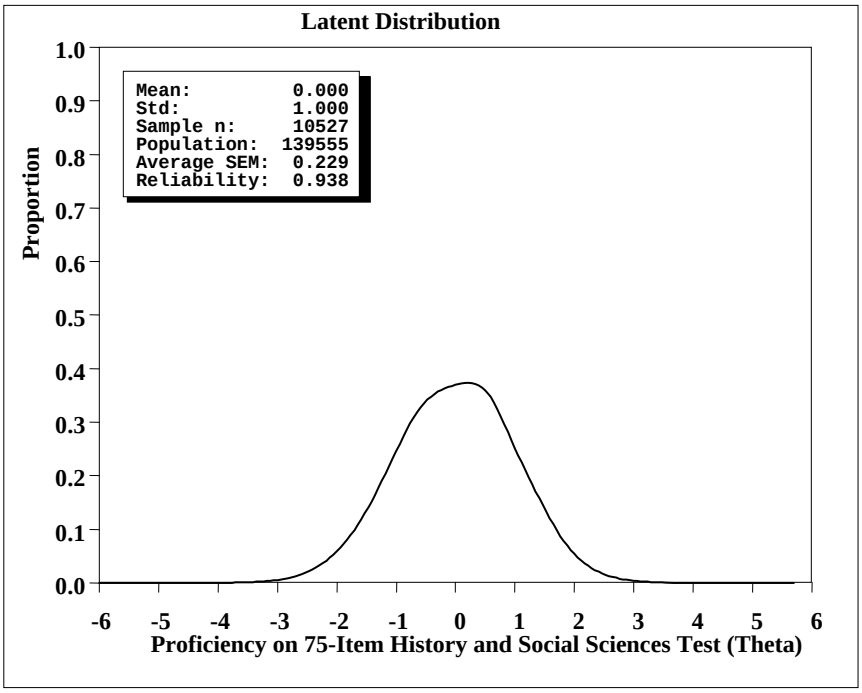


Figure 13: Latent Ability Distribution for 75-Item History and Social Sciences Test

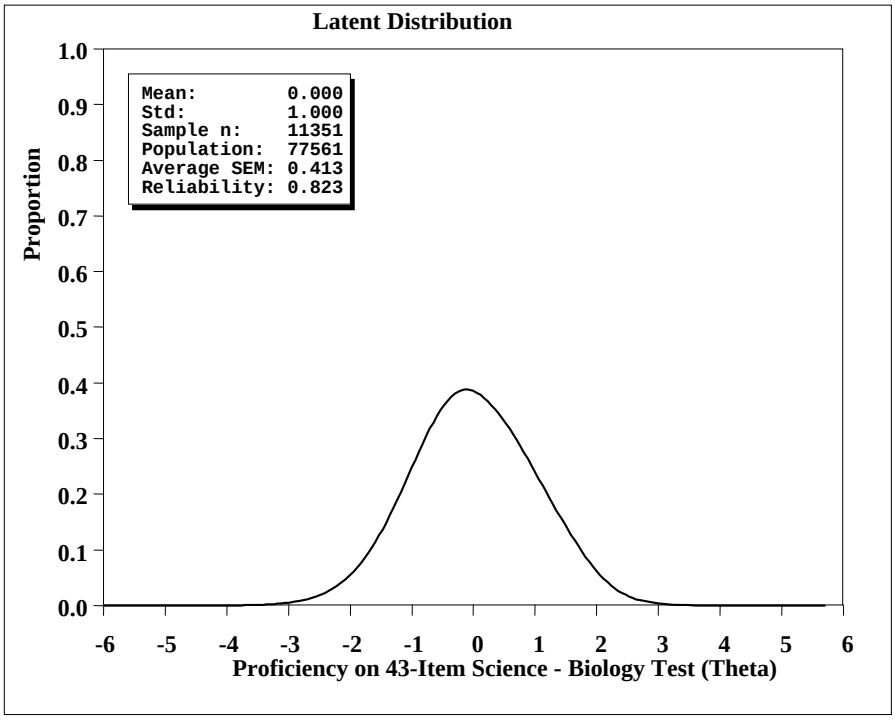


Figure 14: Latent Ability Distribution for 43-Item Science (Biology) Test

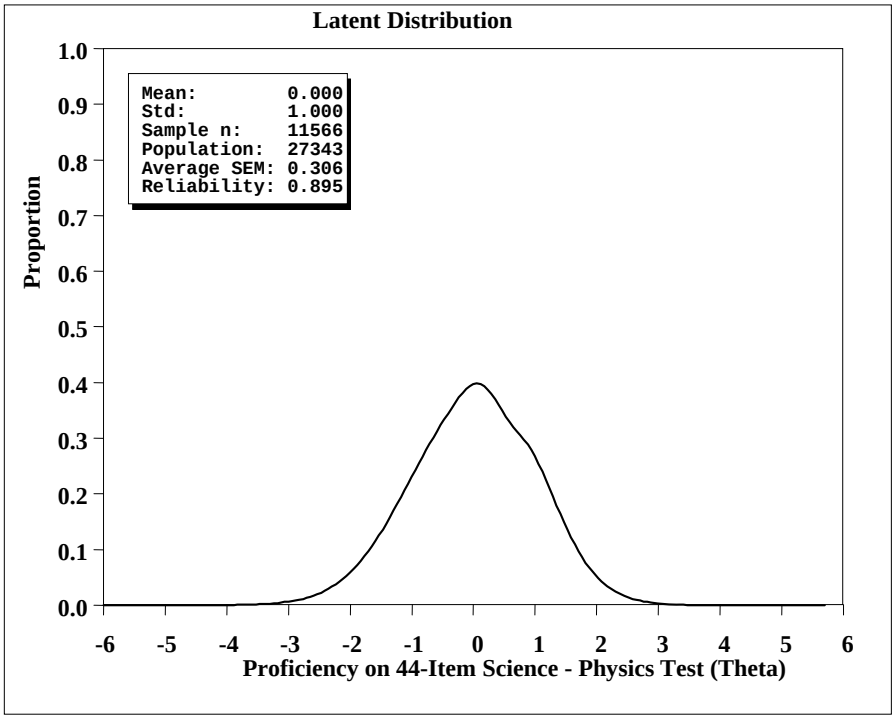


Figure 15: Latent Ability Distribution for 44-Item Science (Physics) Test

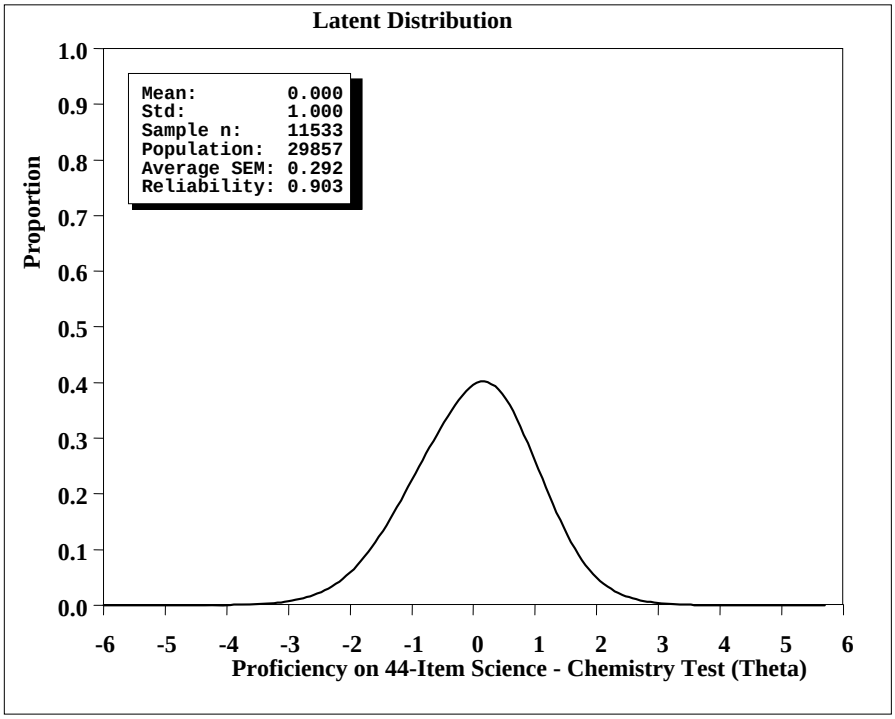


Figure 16: Latent Ability Distribution for 44-Item Science (Chemistry) Test

Test characteristic curve (TCC)

The test characteristic curve (TCC) is a model-based estimate of the number-right raw score as a function of examinees' ability and estimated item parameters. The TCC graphs show the relationship between the model-based raw score (typically known as IRT true score or domain scores) and the ability (proficiency) estimate. IRT ability estimates are linear, while raw scores are non-linear. The TCC is obtained by summing the probability of a correct response at a given ability level across all of the test items. The probability of a correct response across all ability levels is reported in the item characteristic curve (ICC) graphs. The slope of the TCC is related to test information.

Figure 17 shows the TCC for PSU Mathematics test. Examinees at 1.0 on the proficiency scale (x-axis) have an expected PSU Mathematics observed score of about 52 raw points or about 70 percent correct score. Students at 2.0 on the proficiency scale have an expected PSU Mathematics observed score of about 70 raw points or about 95 percent correct score. The figure also shows that the TCC steepness takes place over the ability range from 0 to +2.0, the steeper the slope the greater the test information. The TCC seems reasonable for a test composed of 74 multiple-choice items that were drawn to spread examinees' test scores as in norm-referenced type of interpretations.

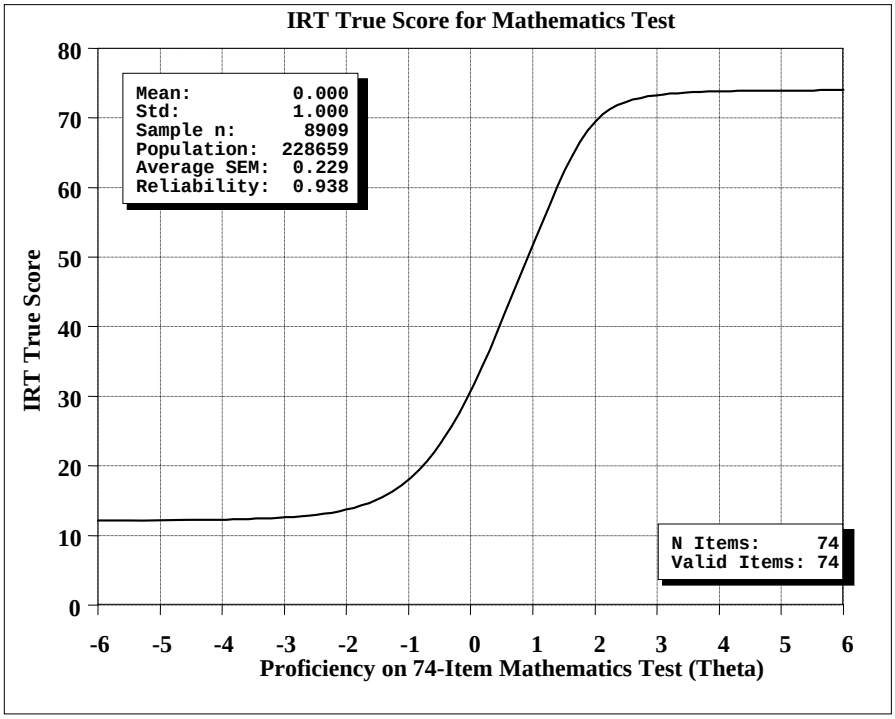


Figure 17: Test Characteristic Curve for 74-Item Mathematics Test

Figure 18 through Figure 23 show test characteristics curves for PSU Language and Communication, History and Social Sciences, Biology, Physics, and Chemistry tests, respectively. Characteristics of the TCC are comparable to those characteristics highlighted for the Mathematics TCC. Collectively, they show an intention of test developers to spread examinees' test scores out in a way that is expected for a norm-referenced type of interpretation. Interestingly, TCC for PSU Language and Communication showed the least steepness among the TCC.

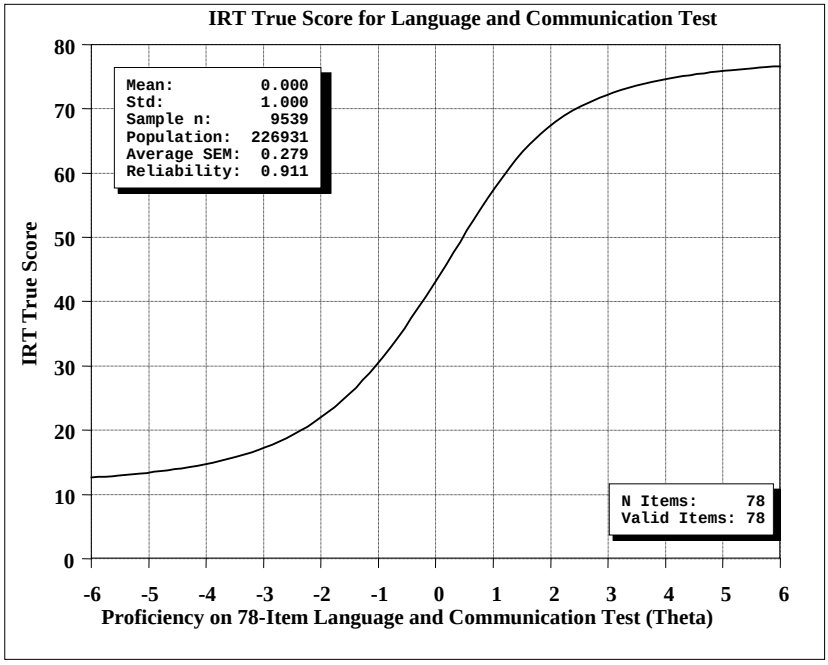


Figure 18: Test Characteristic Curve for 78-Item Language and Communication Test

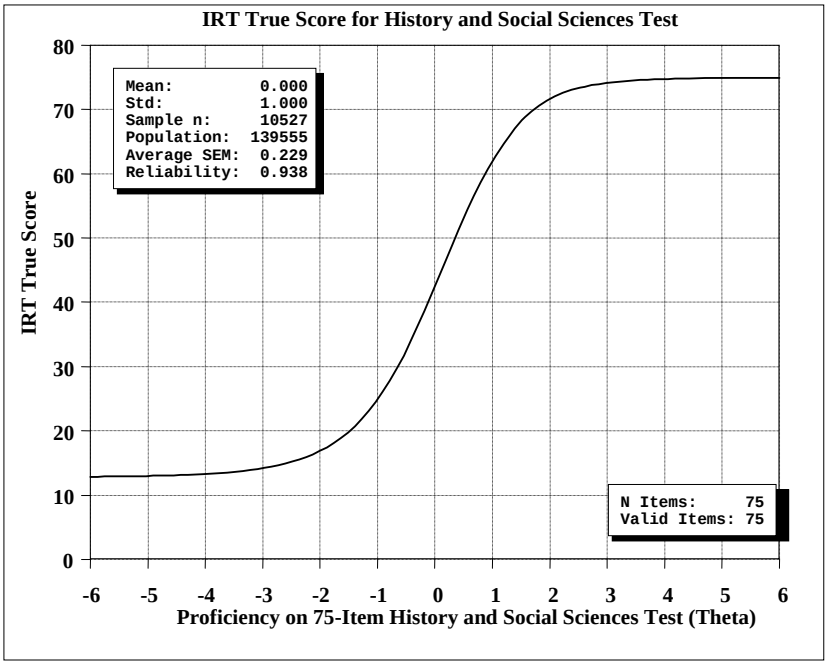


Figure 19: Test Characteristic Curve for 75-Item History and Social Sciences Test

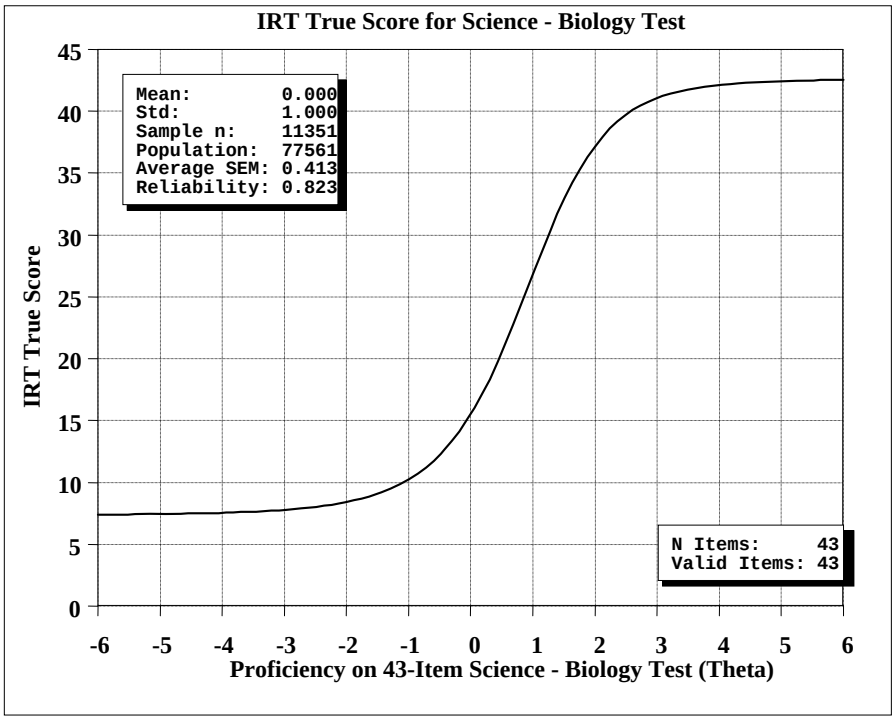


Figure 20: Test Characteristic Curve for 43-Item Science (Biology) Test

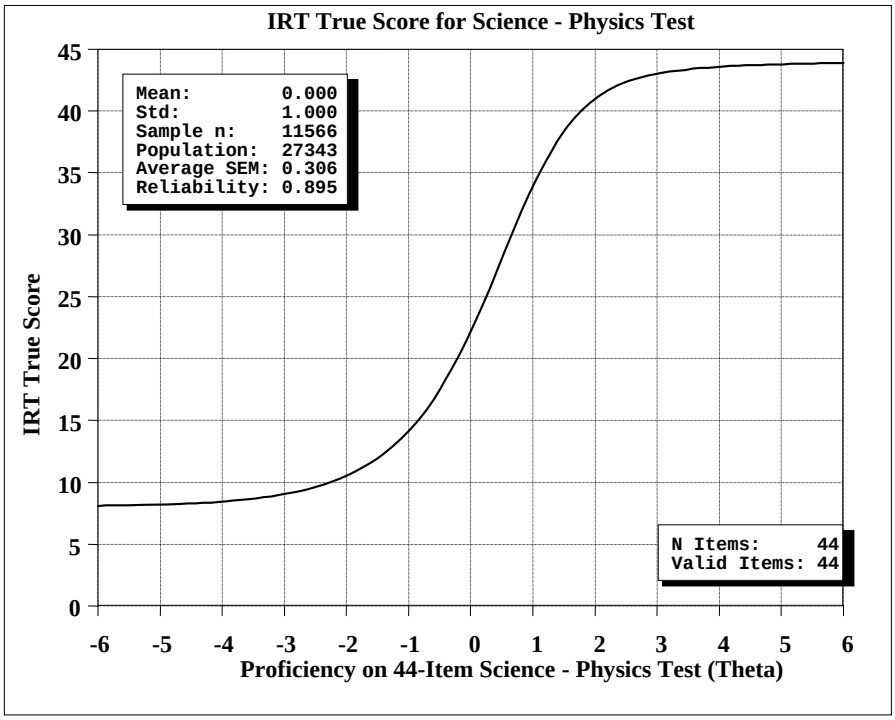


Figure 21: Test Characteristic Curve for 44-Item Science (Physics) Test

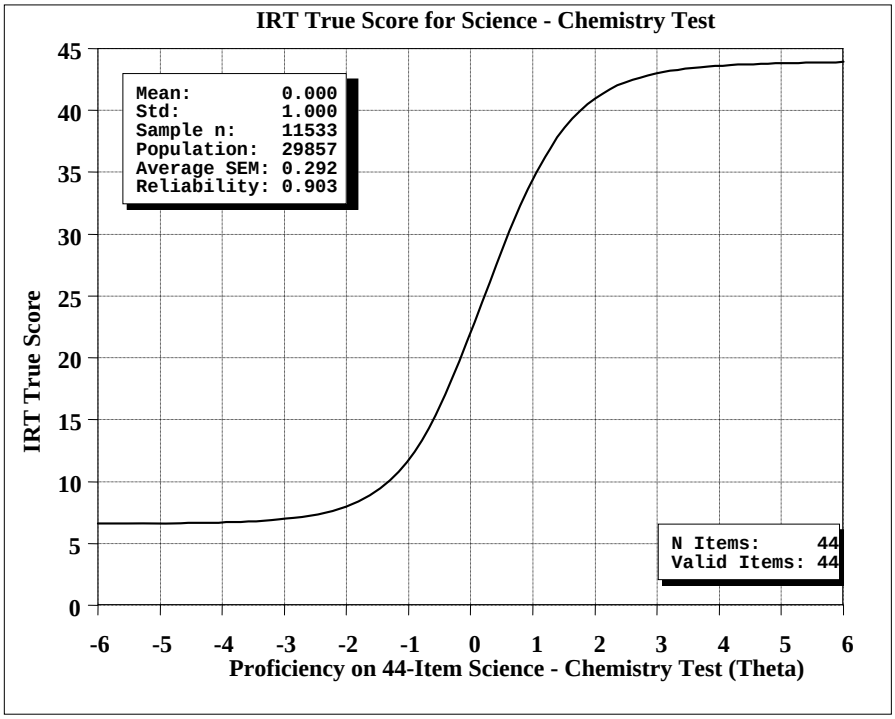


Figure 22: Test Characteristic Curve for 44-Item Science (Chemistry) Test

Test information function (TIF)

Figure 23 shows the TIF for PSU Mathematics test. An important characteristic of the TIF is the additive nature of the contribution of each item to the total information (Hambleton & Swaminathan, 1985). The test information ranges from about 0 to 60 points, indicating that Mathematics items are playing a fairly good role in contributing pieces of independent information to the total test information.

The test achieves its maximum amount of information between 1.0 and 2.0 ability points. It is at this region where students with percent correct scores between 70% and 95% belong. Also, it is at this region where admission decisions are made for Chile's most demanding careers. Whereas an ability score of 1.0 correlates with 610 scale points, an ability score of 2.0 correlates with 720 points. Interestingly, the precision of the test also seems to be adequate for the range between 0.0 and 1.0 ability points, where students with percent correct scores between 41% and 70% belong. It is also the region where decisions are made by Chile's less-demanding careers.

Interestingly, test information decreases between 2.0 and 3.0 proficiency score points. However, the great majority of applicants show proficiency scores at or below 2.0. The TIF seems reasonable for a test composed of 74 multiple-choice items. However, when considering the norm-referenced nature of the PSU, the test falls short of providing equal measurement precision along the proficiency score. It is relevant to state that the amount of test information, although variable across the ability score, highlights ability regions where universities likely set their admission cuts.

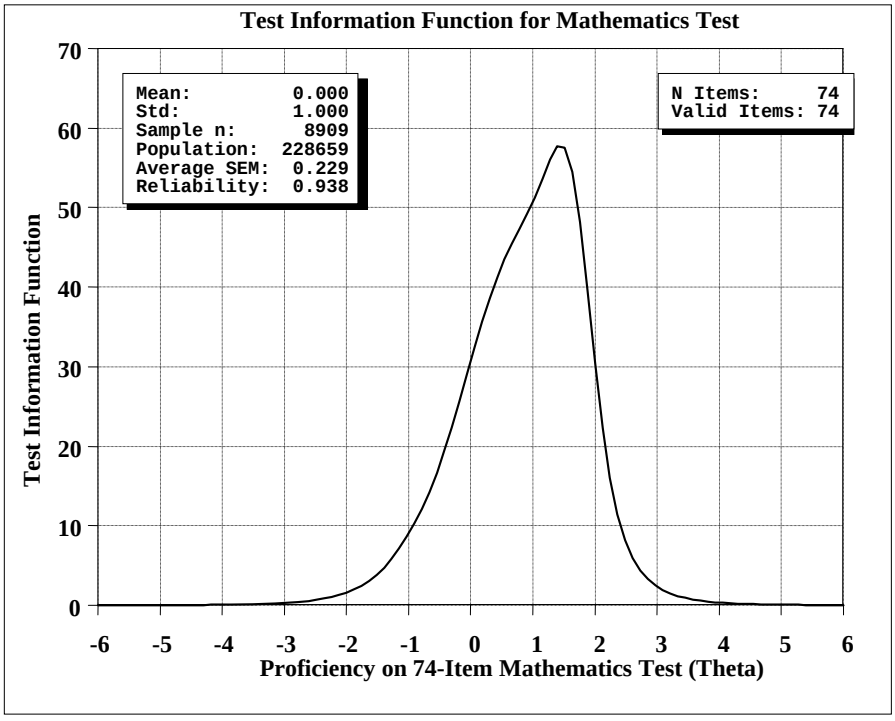


Figure 23: Test Information Function for 74-Item Mathematics Test

Figure 24 through Figure 28 show test information functions (TIF) for PSU Language and Communication, History and Social Sciences, Biology, Physics, and Chemistry tests, respectively. For Language and Communication, the test information ranges between about 0 and 18 points. The attained test information is low for a test composed of 78 items, which indicates low discrimination values of the reading items. This issue was initially spotted from Figure 18, in which the TCC showed a low degree of steepness. Collectively, test construction efforts followed for this test show the intention to spread applicants' test scores out as in norm-referenced contexts. The reliance on CTT framework to guide the test construction activities has hindered the development (a quality control check) of a PSU Language and Communication test with high levels of test information. Additionally, the quality of the items (particularly the degree of discrimination) may have been sub-optimal and contributed small pieces of information to the total test information.

It is relevant to state that the amount of test information for PSU Language and Communication, although less than desirable, has its picks at ability regions where universities likely set their admission cuts.

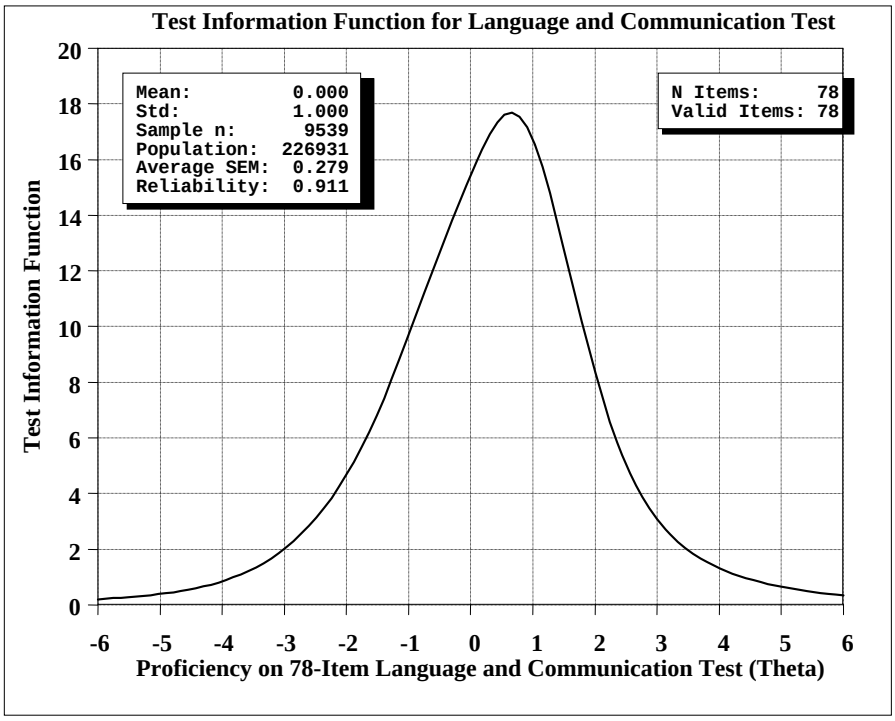


Figure 24: Test Information Function for 78-Item Language and Communication Test

For the History and Social Sciences test, the test information ranges between about 0 and 34 points. The attained test information is moderate for a 75–item test. The quality of several of the items (particularly the degree of discrimination) may have marginally contributed to the information function for the total test. In addition, the test seems to serve better in providing information to universities that set their admission cuts at the middle of the ability distribution. The test may serve marginally to those universities setting admission cuts at higher levels.

It is relevant to state that the amount of test information for the History and Social Science test, although marginal, highlights those ability regions where social science careers may be likely to set their admission cuts.

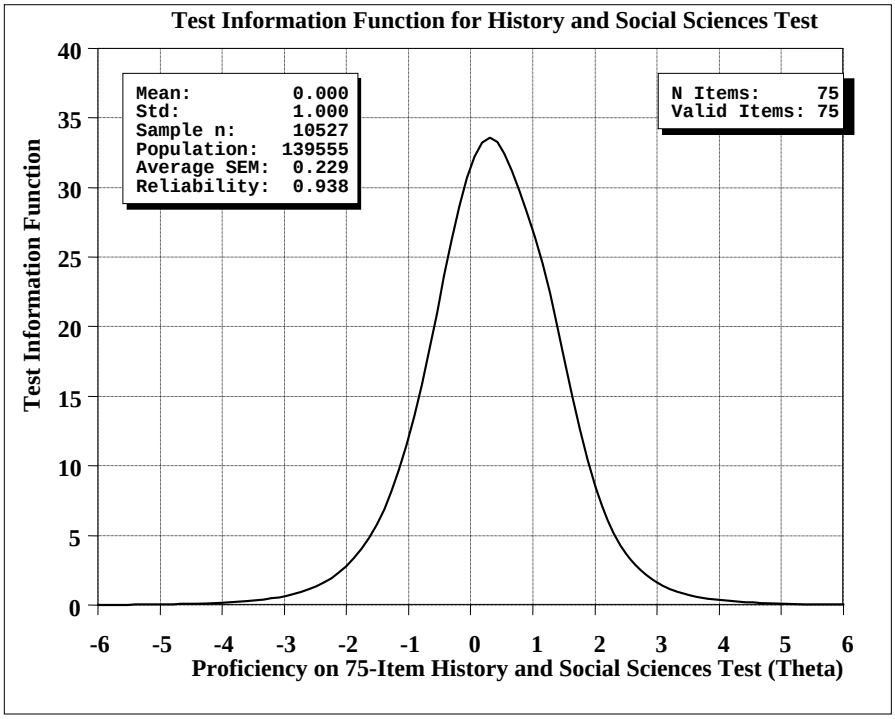


Figure 25: Test Information Function for 75-Item History and Social Sciences Test

Figure 26 through Figure 28 show the TIF for the PSU Science test. The test information ranges between 0 to about 24 points indicating that Science items are playing a fairly good role in contributing pieces of independent information to the total test information. Whereas the Biology and Physics tests achieve their maximum amount of information at about 1.0 ability point (e.g., 610 scale score points), the Chemistry test achieves its maximum information at 0.30 ability point. Because of the above characteristics, the Science tests are targeting information within regions of the ability scores where universities are likely making their admission decisions, with the Biology and Physics tests providing the most amount of information.

Interestingly, test information decreases between 2.0 and 3.0 proficiency score points. However, the great majority of applicants show proficiency scores at or below 2.0. The TIF

seems reasonable for the Science tests (Biology, Physics and Chemistry) composed of 44 multiple-choice items.

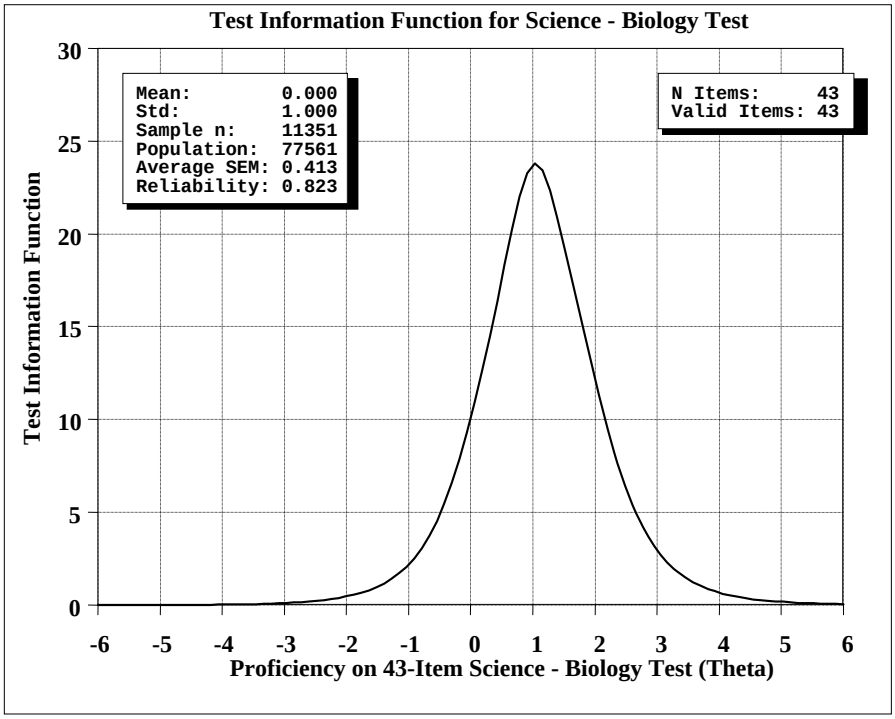


Figure 26: Test Information Function for 43-Item Science (Biology) Test

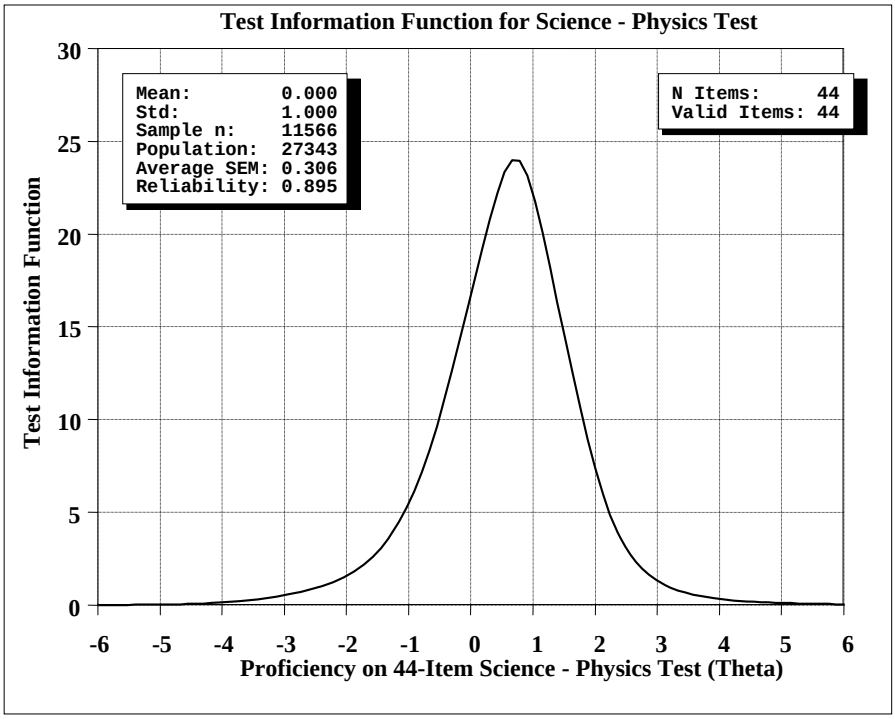


Figure 27: Test Information Function for 44-Item Science (Physics) Test

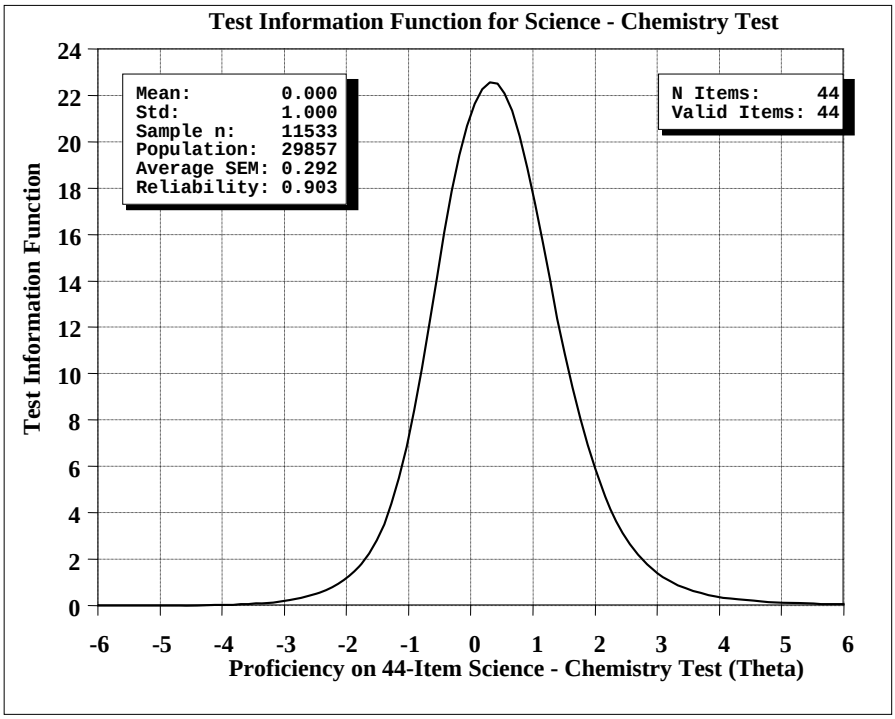


Figure 28: Test Information Function for 44-Item Science (Chemistry) Test

Conditional Standard Error of Measurement (CSEM)

The test information function described in the previous section relates inversely to the size of the conditional standard error of measurement presented in this section. That is, the amount of information in a test at an ability level is inversely related to the measurement error of ability estimates at that point. The section provides a high level summary of the conditional standard error.

The standard error of measurement conditioned at a given level of proficiency is known as the conditional standard error of measurement (CSEM), and it is the reciprocal of the square root of the TIF at that same level of proficiency. The CSEM graph shows the expected measurement error across all ability levels. When comparing the graph of the CSEM to the graph of the TIF, it is important to remember that the vertical axes of the two graphs are on two different scales. Despite the difference in these scales, the two graphs provide comparable results.

The graph of the CSEM tends to be smaller at the middle point of the proficiency distribution and to be larger at the upper and lower tails of the proficiency distributions. These characteristics are often observed for tests targeting applicants with samples of items with appropriate difficulty. That is, because test information should be concentrated at the point on ability distribution where decisions are made, measurement error will generally be lowest at those points. CSEM increases in the tails of the latent ability distribution, where there are relatively few items and few people contributing with information. Note that the values of CSEMs are in the same units of measurement as those for the proficiency scores. CSEM values can be converted to the IRT score scale in the lookup table above (Table 87: IRT Ability Scale and Scale Score (Mean=500 and SD=110)) by multiplying the CSEM by 110, which is the deviation of the scale score transformation.

Figure 29: shows CSEM for PSU Mathematics test. The test achieves minimal amount of measurement error in the 1.0 to 2.0 range of proficiency scores. Analogous to the test information function findings, it is at this region where students with percent correct scores between 70% and 95% belong. Interestingly, the conditional standard error of measurement increases more rapidly for ability scores greater than 2.0 than for ability scores smaller than 1.0. These findings are in accordance with TIF findings that show larger amounts of measurement precision of scores falling within the region where universities are likely to make their admission decisions. The CSEM seems reasonable for a test composed of 74 multiple-choice items.

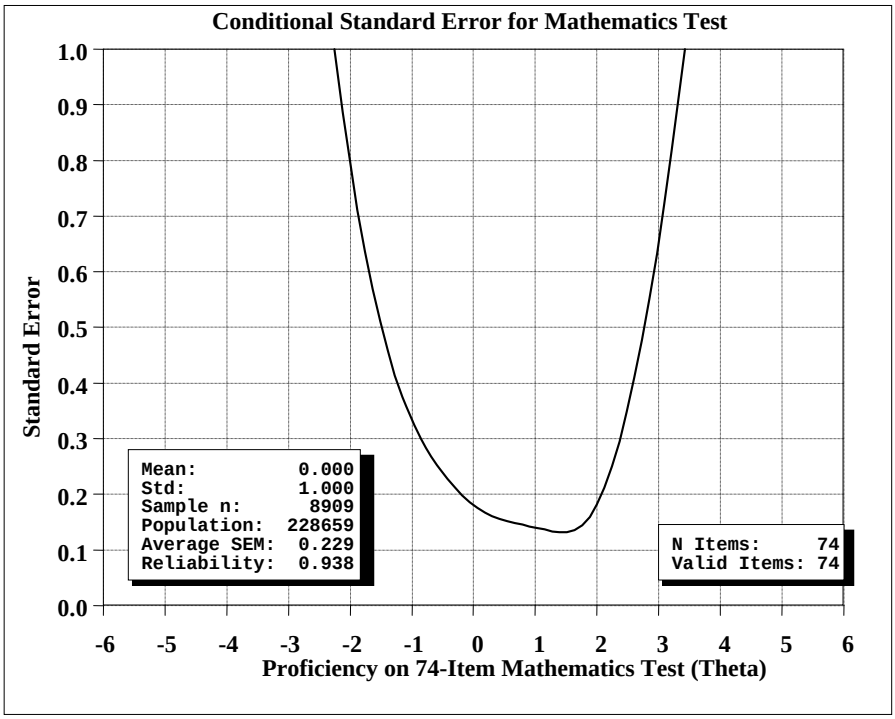


Figure 29: Conditional Standard Error of Measurement for 74-Item Mathematics Test

Figure 30 through Figure 34 show conditional standard error of measurement (CSEM) for PSU Language and Communication, History and Social Sciences, Biology, Physics, and Chemistry tests, respectively. Amounts of CSEM are inversely related to the amount of information functions attained by the tests. It can be seen from the figures that levels of measurement error are larger for the Language and Communication test, followed by the History and Social Sciences test. These two tests are of comparable length to the Mathematics test.

The amount of conditional standard error of measurement that the Science tests achieve translates into a high level of measurement precision. According to the test information functions observed for the Science tests, the size of measurement error was smaller at the ability levels with the highest information. Larger amounts of measurement precision fall within the region where universities likely make their admission decisions. Collectively, the Science tests show an element of intentionality on the reduction of measurement error at regions where Chilean universities set their admission cuts.

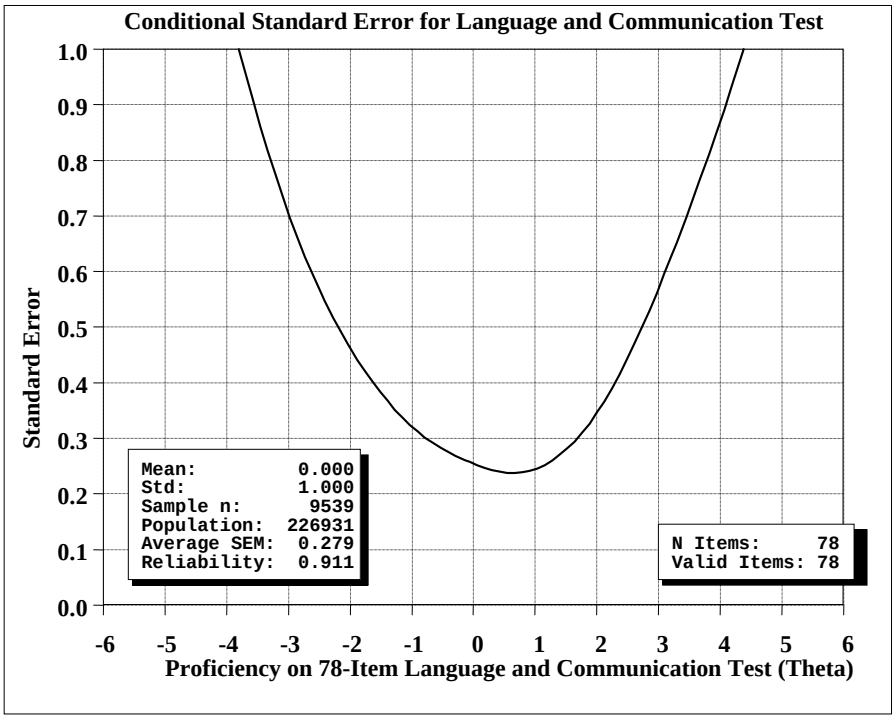


Figure 30: Conditional Standard Error of Measurement for 78-Item Language and Communication Test

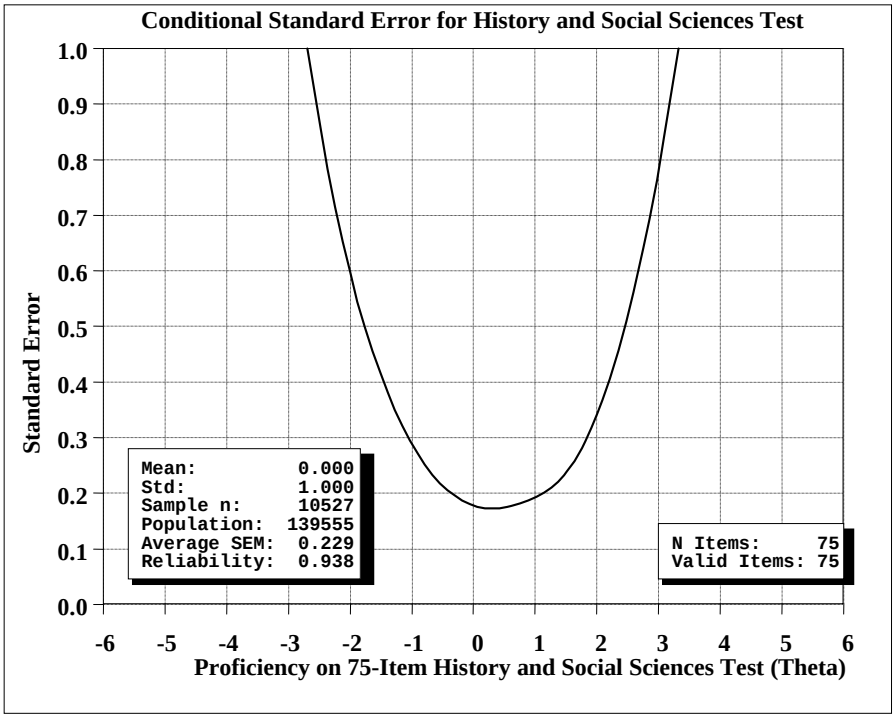


Figure 31: Conditional Standard Error of Measurement for 75-Item History and Social Sciences Test

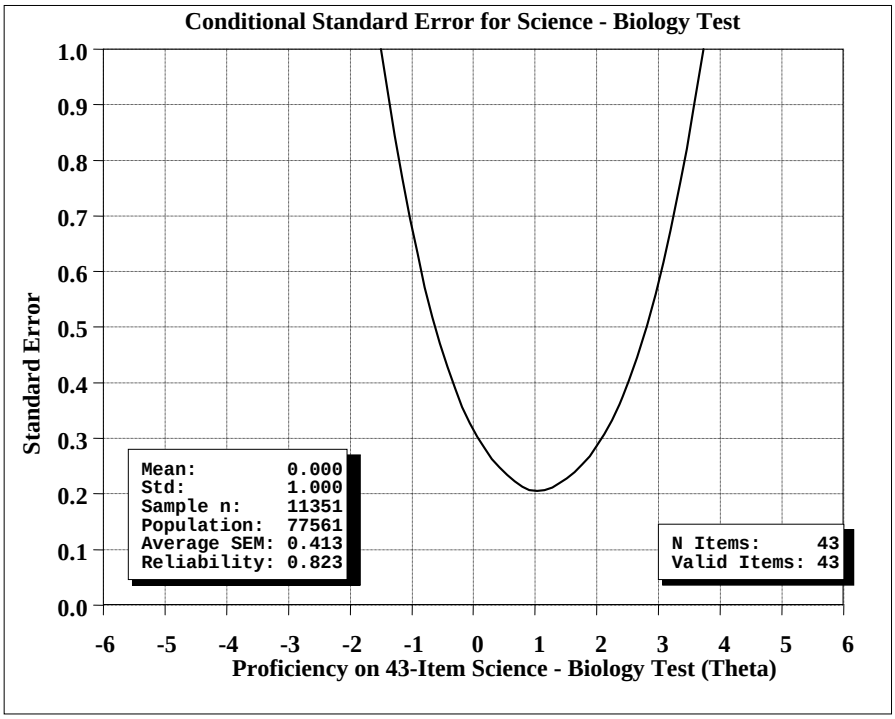


Figure 32: Conditional Standard Error of Measurement for 43-Item Science (Biology) Test

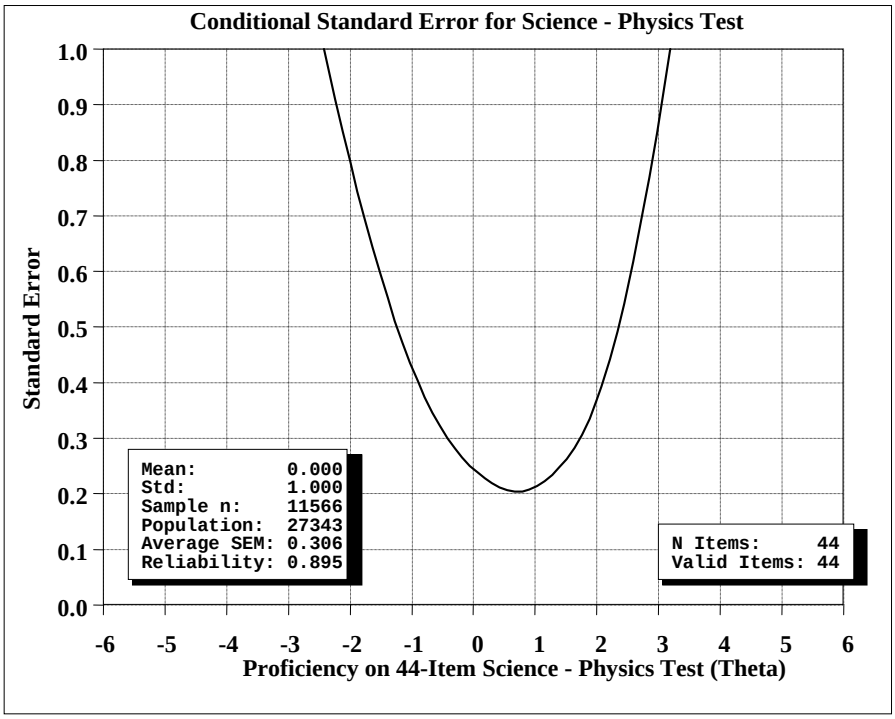


Figure 33: Conditional Standard Error of Measurement for 44-Item Science (Physics) Test

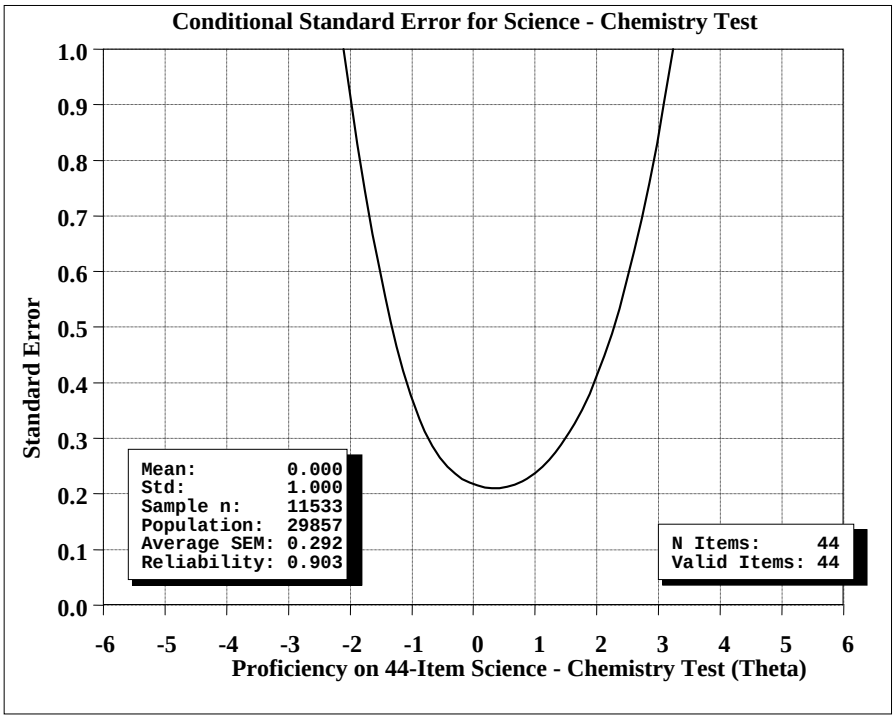


Figure 34: Conditional Standard Error of Measurement for 44-Item Science (Chemistry) Test

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE (2010). *Studio de Confiabilidad de las pruebas de selección universitaria*. Admisión del 2010. Santiago Chile: Autor.

DEMRE (2011). *Studio de Confiabilidad de las pruebas de selección universitaria*. Admisión del 2011. Santiago Chile: Autor.

Frary, R. et al. (1977). Random guessing, correction for guessing, and reliability of multiple-choice test scores. *Journal of Experimental Education*, *46* (1), 11-15.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Norwell MA: Kluwer Academic Press.

Huynh, H. (1976). Statistical Consideration of mastery scores. *Psychometrika, 41*, 65-78.

Livingston, S. & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, *32*, 179-197.

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*, 265-276.

Objective 1.1.j. Propose a model for dealing with cut off points for social benefits, from the perspective of the Classical Test Theory (CTT) as well as from the Item Response Theory (IRT)

First and foremost, PSU test scores produced by DEMRE have the primary purpose of contributing to the postulation scores that are used for admission into university programs. Nevertheless, PSU test scores have also been used by MINEDUC to grant scholarships to incoming university students who qualify to received such awards, a process which lies entirely outside of DEMRE (MINEDUC, 2012). The evaluation team did not receive documentation on the process followed by MINEDUC to define cut scores for granting such social benefits to students. To provide some guidance on this matter, the evaluation team has been asked to propose an approach to set cuts for social benefits utilizing facets and elements shown in Table 88, which were defined during the goal-clarification meeting.

Table 88: Facets and Elements for Recommending a Model to Derive Cut Scores for Assigning Social Benefits

FACETS	ELEMENTS
1. Choice of standard setting method	• Rationale behind the proposed approach • Types of social benefits • Panelists • Information on students with awarded scholarships (college academic performance and attrition rates. • Career information (number of admission places, financial resources available to provide scholarships, number of qualified applicants seeking for scholarships). • Process to calculate cut score precision and classification consistency rates.
2. Choice and training of standard setting facilitators & panelists	• Demographically diverse • Professionally diverse • Geographically diverse • Thorough understanding of current curriculum • Thorough understanding of current students • Training to provide thorough understanding of standard setting procedure
3. Current & past PSU cut scores & pass rates	• Rationale for current and past standards • Changes in standards over time • Fairness issues for teachers • Fairness issues for students
4. Policy considerations	• Ministry needs for public consumption • Demonstration of improvement • Appropriate representation of actual student ability • Who has final authority to set cuts?

5. Inputs into standard setting	• Previous/current cut scores & pass rates • Test or item data • Impact data
6. Performance level descriptors	• Number of performance level descriptors • Level of detail of descriptors • Overlap between levels • Examples given for each
7. Documentation of process	• Rationale for process used • Explanation of process • Description of participants • Explanation of final results • Expectations for implementation

The evaluation team has extensive experience in designing cut score derivation activities using CTT and IRT approaches. In our experience, the specific methodology chosen for setting cut scores should be determined after carefully detailing the context of the program. With our current understanding of the PSU context, the evaluation team engaged in reviewing relevant literature on standard setting that takes into account the primary use and interpretation of PSU scores and balances it with policy considerations and social consequences, such as first year university grade point average and the graduation rate. Typical methods for setting cut scores used in education focus on expert judgments about content mastery. However, identification of content mastery may not be sufficient to meet the policy goals of the PSU test. For example, if one of the stated goals of the PSU is to rank order candidates to fill a limited number of university openings each year, the relative mastery of the accepted candidates may vary from year to year.

Given the stated goal of this facet, the evaluation team proposes an approach that considers domain mastery as one foundational layer and social consequences as another foundational layer. The first layer drives to identify the level of knowledge, skills and abilities defined by a blue ribbon panel of university professors and MINEDUC policymakers and to translate this definition into a PSU admission score. The second layer focuses on taking into account policy considerations, social consequences and historical data to fine-tune the cut score. The second layer uses a reaction panel of MINEDUC policymakers to review historical data on the number of scholarships available each year as well as the number of students receiving scholarships, the students' performance, their attrition rate, and their graduation rate.

Our recommendations for a process to set a cut score for awarding scholarships to incoming university students cover areas that need attention for all standard settings, as described in Cizek (2007):

- Purpose
- Choosing a standard setting method
- Performance level labels
- Performance level descriptions
- Key conceptualizations
- Selecting and training standard setting participants
- Providing feedback to panelists

The following subsection describes a proposed approach to set cut scores for assigning social benefits in response to Objective 1.1.j.

GENERAL DESCRIPTION

For a typical university admissions process in which there are more students applying for scholarships than the available number of scholarships, some kind of applicant selection must be performed utilizing sound approaches. The aim of the process to grant scholarships is to rank order the applicants seeking social benefits according to some model that takes into account the following facets: (1) sound methodology, (2) characteristics of panelists and facilitators, (3) policy considerations, (4) PSU data and applicants characteristics, (5) impact data and historical academic performance in college and (6) documentation of process and results.

A commonly and generally accepted principle for guiding the award of scholarships is to grant scholarships those applicants who are most likely to succeed in careers of their choice. School-related academic achievement, test scores on college entrance examinations, and applicants' financial needs-based assessments are among the most common criteria gearing processes to grant scholarships. In the United States, scholarships are merit-based awards designed to help defray the cost of a college education. A large amount of money is available from a wide array of organizations including private companies, non-profit organizations and universities.

In Chile, scholarship money comes from the national budget. The process for granting scholarships takes into account the applicants' PSU test scores among other variables of interest. MINEDUC has defined the general requirements for applying to scholarships. One portion of the requirements pertains to important dates for applying for a scholarship. The other portion outlines conditions governing the use of PSU test scores. See, for example,

> The PSU score obtained in the admissions process for 2011 or 2012 was used for scholarships and the University Credit Solidarity Fund for the career selected by the applicant. The Academic Excellence and PSU Points Scholarship considered only the PSU scores obtained during the 2012 admissions process. (Donde Quede, 2012, 13 Octubre)

There is a wide array of scholarships available to students pursuing post-secondary education. More information on the types of scholarships and qualifications are available from MINEDUC's Internet portals www.becasycreditos.cl.

Table 89 shows PSU scores reported explicitly by Chile's scholarship program. The eligibility requirement is presented by the type of higher education scholarship available in 2012.

Table 89: Chile's Higher Education Scholarship Programs

Scholarship	High school academic performance	PSU test score	Socio-economic standing
Beca de Excelencia Académica	At the top 5% of high school GPA (municipal and subsidized). Applicants awarded with *Beca Propedéutico* will be exempted from GPA requirement.	Depends on admission criterion followed by the accredited higher education institution and career	Belonging to the first four quintiles of SES in the nation
Beca Bicentenario	Not specified	At least 550 points (Language and Communication and Mathematics)	Belonging to the first two quintiles of SES in the nation
Beca Puntaje PSU	Not specified	National or Regional PSU test scores	Belonging to the first four quintiles of SES in the nation
Beca Juan Gómez Millas	Not specified	At least 550 points (Language and Communication and Mathematics)	Belonging to the first two quintiles of SES in the nation
Beca Juan Gómez Millas Extranjeros	Depending on criterion followed by the accredited higher education institution and career	Depending on criterion followed by the accredited higher education institution and career	Belonging to the first two quintiles of SES in the nation
Beca Vocación de Profesor -Pedagogía	For applicants whose PSU is at least 580 points, to qualify for a level 1 scholarship they must be at the top 5% high school GPA (municipal and subsidized).	PSU score according to amount of scholarship: Level 1: 600 points, Level 2: equal or greater than 700 points; Level 3: equal or greater than 720 points. (Language and Communication and Mathematics)	SES is not considered.
Beca Hijo Profesionales de la Educación	High school GPA of at least 5.5	At least 500 points (Language and Communication and Mathematics)	Belonging to the first four quintiles of SES in the nation. The SES requirement is exempted for 100 best applicants with at least 600 PSU points and high school GPA of at least 6.0 points.

Beca Vocación de Profesor -Licenciatura	Not specified	PSU score according to amount of scholarship: Level 1: 600 points; Level 2: equal or greater than 700 points. (Language y Communication y Mathematics)	SES is not considered.
Beca Propedéutico		Depends on admission criterion followed by the accredited higher education institution and career	Belonging to the first two quintiles of SES in the nation

RECOMMENDATION

The approach we would like to recommend makes use of both domain mastery and social consequences. The former aims to identify the level of knowledge, skills, and abilities defined by a panel of university professors and MINEDUC policymakers and to translate that definition into a PSU cut score to tap on academic merit for an scholarship. The latter takes into account policy considerations, social consequences and historical data to fine-tune the PSU cut score. A focus on social consequences would entail MINEDUC policymakers convening a panel to use such information as the historical data on the number of scholarships available each year, the number of students receiving scholarships, as well as the performance, attrition rate and graduation rate of those receiving scholarships.

Our process for setting a cut score for awarding scholarships is detailed below and is modeled after the relevant elements in the process described in Hambleton and Pitoniak (2006, p. 433-470) and Cizek and Bunch (2007, p. 35-67).

Choosing a standard setting method

The method that we recommend to MINEDUC for setting the cut score for scholarship purposes is the Hofstee method. We recommend this method for several reasons.

First, the Hofstee method is an example of a compromise[9] standard setting method. Hofstee (1983) coined this term when he developed his method to capture the dual nature of standard settings. That is, even criterion-referenced standard setting judgments are tempered by norm-referenced expectations. His method makes explicit use of both criterion- and norm-referenced information to derive cut scores. This is important for awarding scholarships because there are finite resources to distribute and norm-referenced selectivity targets those resources to students in a manner that should be considered along with criterion-referenced subject area mastery.

Second, the Hofstee method is not tied to a specific measurement model. That is, it is a method that works equally well within the framework of either CTT or IRT.

[9] The term "compromise" does not connote the lessening or dilution of either of the two criteria of judgment; rather, it signifies the integration of the two in a manner that is more multifaceted than either criterion applied individually.

Finally, the Hofstee method makes economic use of time. As Cizek and Bunch (2007) state:

> The judgmental task required in the Hofstee method requires the standard setting panelists to answer two questions that examine their criterion-referenced expectations and two questions regarding their norm-referenced expectations of achievement on the test. (p. 210)

The Hofstee method is by no means free from criticism in the standard setting literature (Cizek & Bunch, 2008). Among the two most salient ones are the compromise nature of the method and its low popularity relative to traditional standard setting methods. As a compromise method, the Hofstee approach seeks to balance content and performance standards with political agendas, economic pressures, and policy concerns to reach a solution. In this sense, the selection of a cut score (often a binary one like in pass/fail) will not be made only on a scientific basis.

The other issue facing the Hofstee method is its low popularity in use and research. The popularity of the approach has been proportionally related to the low amount of research carried out for the approach. Unfortunately for the Hofstee method, more traditional approaches have been found more appealing to practitioners and researchers during the past fifteen years. Due to the binary nature of the Hofstee method (such as in pass/fail decisions), researchers have found it to be less attractive than traditional standard setting approaches that allow for multiple cuts.

We have modified the original Hofstee questions and notation to reflect the scholarship for the PSU. The four questions asked of standard setting panelists are as follows:

- "What is the *highest* PSU cut score that would be acceptable, even if *every* student attains that score?" This value is a panelist's estimate of the maximum level of knowledge that should be required of students in order to be awarded a scholarship and is denoted K_{max}.

- "What is the *lowest* PSU cut score that would be acceptable, even if *no* student attains that score?" This value is a panelist's estimate of the minimum level of knowledge that should be required of students in order to be awarded a scholarship and is denoted K_{min}.

- "What is the *maximum* acceptable percentage of students that will *not* be awarded a scholarship?" This value also represents a panelist's judgment of the *lowest* percentage of students who *will* receive a scholarship and (for technical reasons) is denoted P_{max}.

- "What is the *minimum* acceptable percentage of students that will *not* be awarded a scholarship?" This value also represents a panelist's judgment of the highest percentage of students who *will* receive a scholarship and (for technical reasons) is denoted P_{min}.

The first two of these questions represent what would be the upper and lower criterion-referenced expectations regarding performance on the PSU tests for the purposes of being awarded a scholarship. K_{max} would represent a level of achievement on the PSU that is considered so high that, even if every student scored at that level or higher, they would be

worthy of a scholarship. On the other hand, K_{min} would represent the absolute lowest level of achievement on the PSU that could be tolerated, even if no students received scholarships.

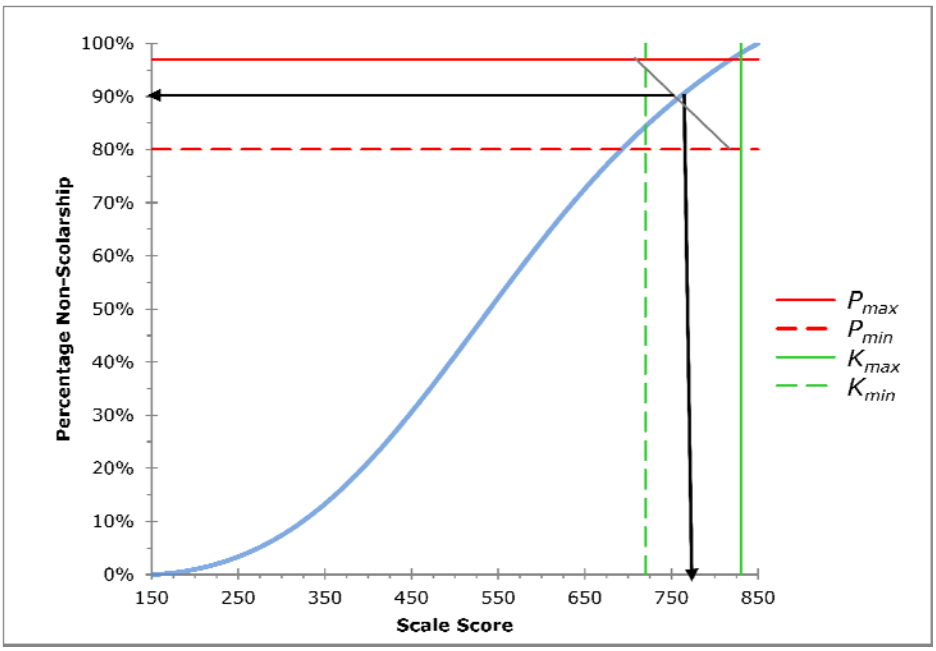


Figure 35: An Example of the Hofstee Cut Score Derivation

The last two questions allow for norm-referenced information to be incorporated into the standard setting judgment. As described above, selectivity is an aspect of scholarship programs that needs to be discussed and is usually thought of in normative terms. If too many students receive a scholarship, then the selection process might be thought of as not being selective enough. This is what the value of P_{max} controls. Conversely, if very few or no students receive a scholarship, then the process would be considered overly selective. This is controlled by the value of P_{min}.

In practice, the standard setting panelists would be given a form to obtain their values of K_{max}, K_{min}, P_{max}, and P_{min}. Each of the four sets of values would be averaged to arrive at values representing the group's criterion-referenced and norm-referenced expectations.

Figure 35 details a hypothetical application of the Hofstee method[10]. Assume for this example that the values of K_{max}, K_{min}, P_{max}, and P_{min} are approximately 830, 720, 97%, and 80%. These four values have been plotted on the graph with the green vertical solid and dashed lines representing K_{max} and K_{min}, and the red horizontal solid and dashed lines representing P_{max} and P_{min}. A grey line segment has been added to the graph from the point (K_{min}, P_{max}) at the upper-left side to the point (K_{max}, P_{min}) at the lower-right side. As Cizek and Bunch (2007) state, "According to Hofstee, it is along this line that all 'compromise' solutions to the cut score may be found" (p. 211).

[10] Note that this example just uses the context of the PSU test for illustrative purposes and does not imply the use of actual PSU test data or an endorsement by the evaluation team for a given cut score or other values.

In addition to the lines and points described above, the observed test score distribution for the PSU is plotted in blue on the graph. Under the Hofstee method, the intersection of this curve with the previous line represents the optimal compromise cut score subject to the constraints given by K_{max}, K_{min}, P_{max}, and P_{min}. The observed test score distribution is presented as a cumulative distribution function. For each point on the horizontal axis corresponding to a PSU scale score, there is related a point on the vertical axis corresponding to the percent of students that would not receive a scholarship if that PSU scale score was used as a cut score. As the PSU scale score increases, more and more students would be below the cut score, and fewer students would be awarded a scholarship.

The final step in the process is to project the intersection of the line segment from (K_{min}, P_{max}) to (K_{max}, P_{min}) and the observed test score distribution onto axes representing the PSU scale scores and the percentage of students not receiving a scholarship. For this example, this intersection point would correspond to a cut score of approximately 775 and a percentage equal to 90% not receiving a scholarship or 10% receiving a scholarship.

Choosing Panelists

Under the assumption that the training, content familiarization, and review of key conceptualizations can be accomplished in approximately a one half-day session, the remaining of the Hofstee method—that is, the completion of the actual judgmental tasks required of the participants—ordinarily requires less than one additional hour, making the Hofstee method one of the most time-efficient ways of setting cut scores (Cizek & Bunch, 2007, p. 209-210).

Panelists should be familiar with higher education demands and incoming entry level university students. A broad representation of college faculty from geographical regions and higher education institutions is desirable. Of particular interest is achieving representation of universities and higher education institutes from the country. It is desirable to involve a minimum of 20 panelists to achieve smaller variability around proposed cut scores.

Policy Consideration

The step of defining the policy surrounding the standard setting is crucial to the successful application of the PSU for awarding scholarships. The more clearly and explicitly MINEDUC does this, the more likely it will succeed in its aim.

For example, it is important that MINEDUC considers the basis upon which the scholarship for students is being judged. Thus, should scholarships be awarded for high achievement in individual subjects such as Mathematics or Biology, or should they be awarded for high overall achievement across several subjects? If scholarships are to be awarded for each individual subject, it may be necessary to set a separate cut score for each subject. This would require the use of several standard setting sessions and/or committees of standard setting panelists. However, if overall achievement is to be awarded, a less expensive process may be used to create a composite PSU score across two or more subjects, which would entail using only a single standard setting session and committee to set a single cut score. A logical possibility for rewarding general overall achievement might be to create a composite score based on the PSU Mathematics and Language scores, as all examinees are required to take the tests in these subjects.

It is also important for MINEDUC to consider how different sources of information should be used in making standard setting judgments. More specifically, should the scholarships be awarded solely on the basis of PSU test scores, or should they take into account other sources of information? If part of the purpose in awarding scholarships is to encourage

students to do well on the PSU, then MINEDUC may want to consider cut scores that only take into consideration PSU scores. On the other hand, a richer indication of academic achievement is more likely if other sources of information are taken into account. This would require a composite score to be formed as is currently done with students' NMREs when applying to a university.

Carefully considering issues such as these will enable MINEDUC to articulate a clear purpose for standard setting that the general public can understand and more readily accept.

Data Sources and Impact Data

Identifying data sources for standard setting activities is a key component for building defensibility of cut scores. Cut scores arrived with a lack of understanding of intended populations and some type of performance data may be less defensible than outcomes from standard setting involving relevant data and clear definition of expectations. The evaluation team recommends the following list of information to be made available for standard setting activities:

- Approval rates and cut scores (previous / current)
- Item or test information
- Data to assess the impact of accepting a given cutoff score
- Descriptor for the performance level of the candidate obtaining the scholarship
- Examples of performance descriptor

Reports

Standard setting documentation provides a line of evidence supporting the validity of proposed cuts. Typical standard setting reports summarize characteristics of the methods, participants, training and data provided to panelists. Additionally, these reports also show impact data showing the distribution of subpopulations receiving scholarships based on proposed cut scores. This source of information provides the technical defensibility of results while gathering evidence of the relative fairness of the proposed cuts. A give cut showing a disproportionate number of male applicants receiving scholarships in relation to female applicants would be seen as suspicious. A final element to be included in the standard setting report is a summary of the panelists' evaluation of standard setting activities. Well-run standard settings meetings receive among others (1) high marks on panelist readiness, (2) understanding of directions, (3) opportunity to interact and present ideas, and (4) satisfaction with proposed cut scores.

Final Remarks

As we stated above, we recommend an approach to dealing with cut scores that considers both domain mastery and social consequences. We believe that the Hofstee method presented here provides MINEDUC with the standard setting approach for accomplishing that goal that is best from a practical and defensible policy point of view.

Although the approach is not exempt from criticism, as any other standard setting process, we believe that setting a cut score for granting scholarships requires a blend of considerations: some of them scholastic, some monetary, and some social. The virtue of the Hofstee approach relies on its capacity for bringing all these perspectives into account to facilitate discussions and compromising decisions within a reasonable time frame. The approach is flexible enough to be used with the current approach to scaling the PSU (i.e., using classical test theory) as well as any foreseeable changes to the system (i.e., using item response theory).

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Donde Quede. (2012, 13 Octubre). *Becas y credítos*. Retrieved from http://dondequede.cl/index.php?option=com_content&view=article&id=51&Itemid=70

Hofstee, W. K .B. (1983). The case for compromise in educational selection and grading. In S. B. Andersen & J. S. Helmick (Eds.) *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass.

Ministerio de Educación de Chile (MINEDUC) (2012). *Becas y créditos 2012 para la educación superior*. Santiago Chile: Autor.

Objective 1.2. Analysis of the adequacy of a single score in the Science test and of the procedures to calculate said score, considering that this test includes elective blocks of Biology, Physics and Chemistry

Over the course of this evaluation, the evaluation team has gained deeper understanding on the surrounding context and process to render a single score for PSU Science. In Chile, the PSU Science test comprises a common section (Biology, Physics, and Chemistry content from the first two years of high school) and three elective modules (Biology, Physics, and Chemistry from the last two years of high school). For the administration of the PSU Science, besides taking a common section, the applicants must render one of the three elective modules. By a policy mandated from the CRUCH, one single PSU Science score is developed through the means of a process labeled as "equating."

The purpose of this objective is to analyze adequacy of the PSU Science score and processes to develop such a score. As explained later in this objective, the PSU Science "equating" uses the score from the common module and a set of nodal points to derive the single score for PSU Science. Since the 2005 admission process, DEMRE has operationally adopted the "equating" methodology which was developed by PSU technical advisory committee (CTA).

The evaluation team read the CTA documentation of PSU Science "equating" process and met with DEMRE staff to fine-tune information on the process to calculate the single Science score. The evaluation team met with relevant stakeholders from DEMRE on March 26, 2012, to participate in a demonstration of the PSU Science equating process. The purpose of the demonstration was to gain deeper understanding of PSU Science equating for the following particular aspects:

- Process to equate PSU Science
- Weights estimation
- Conditional standard error of PSU Science equate score
- Process maintenance and evaluation

The demonstration meeting was performed within DEMRE offices following an agreed-upon schedule developed for the interview. The interaction with DEMRE staff took place for four evaluation facets pertaining to the objective agreed to with the TC during the goal clarification meeting in Santiago, Chile, in January 2012.

In addition to the evaluation of the above facets, the evaluation team provided an analysis of pertinence of PSU Science single score. For these analyses, the evaluation team has addressed two fundamental questions: (1) How reasonable is the reporting of a single score? and (2) What alternatives exist? The evaluation team also discussed the issue of dimensionality of test scores.

The following DEMRE staff participated in the interviews:

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction and the attainment of the meeting purpose.

The international evaluation team has relied on professional standards during its appraisal of the merit and worth of the PSU process for deriving a single Science score for the PSU test. A framework for evaluating PSU approaches for deriving single Science score has been developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 4.10

> A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably. In some cases direct evidence of score equivalence may be provided. In other cases, evidence may come from a demonstration that the theoretical assumptions underlying procedures for establishing score comparability have been sufficiently satisfied. The specific rationale and the evidence required will depend in part on the intended uses for which score equivalence is claimed. (p. 57)

Standard 4.11

> When claims of form-to-form equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions. (p. 57)

Standard 4.14

> When score conversions or comparison procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those conversions or comparisons should be clearly defined. (p. 58)

Standard 4.17

> Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported. (p. 59)

This section is organized into two main themes. One theme provides analysis of the pertinence of combining Science scores into a single score. This section is identified as PART I. The other theme provides analyses of "equating" processes to render a single score in Science for PSU. This section is identified as PART II.

GENERAL DESCRIPTION (PART I)

The PSU test battery comprises four standardized instruments: two obligatory tests (Mathematics and Language and Communication) and two optional tests (Science and History and Social Sciences). One of the two optional tests must be included as a selection factor for each university career. In all the tests forming part of the PSU battery, questions are multiple-choice with five alternative answers, only one of which is correct. The total raw scores of each test are corrected discounting from the number of correct answers, one fourth of a point for each incorrect one.

The PSU Science test has two modules: (1) a common and (2) an elective. Applicants take the common module and must take one of the three elective modules (Biology, Physics or Chemistry). Table 90 shows the distribution of items for common and elective sections of the PSU Science test.

Table 90: Distribution of Items for PSU Science test

<table>
<tr><th colspan="3">Common module</th><th>Elective module Biology</th><th>Elective module Physics</th><th>Elective module Chemistry</th></tr>
<tr><td rowspan="3">Biology (18 items tapping content from first and second year of high school)</td><td rowspan="3">Physics (18 items tapping content from first and second year of high school)</td><td rowspan="3">Chemistry (18 items tapping content from first and second year of high school)</td><td>Biology (26 items tapping content from third and fourth year of high school)</td><td></td><td></td></tr>
<tr><td></td><td>Physics (26 items tapping content from third and fourth year of high school)</td><td></td></tr>
<tr><td></td><td></td><td>Chemistry (26 items tapping content from third and fourth year of high school)</td></tr>
</table>

Two patterns are important to highlight from Table 90. One is the portion of Chile's national high school curriculum targeted by the PSU Science test. The common module is 54 items long and equally taps content curriculum on the Biology, Physics, and Chemistry taught during the first two years of high school. The elective modules for Biology, Physics, and Chemistry are each 26 items long, and they tap curriculum content taught during the last two years of high school. The other pattern found in Table 90 is the test length of the PSU Science test. An applicant choosing to take

Biology as an elective module gets a Science booklet with a total of 80 items (54 items from the common module and 26 items from the elected module—Biology, in this example).

EVALUATION (PART I)

Pertinence of a single score for PSU Science test

The evaluation team addressed the analyses of the pertinence of providing a single score for PSU Science test of Biology, Physics and Chemistry by providing answers to the following two questions:

1. How reasonable is it to provide a single Science score?
2. What alternatives are there to replace the current practice of reporting a single score?

In addition, as part of the above analyses the evaluation team provided a rationale for the role of test dimensionality in tests developed from a domain. The evaluation team clarified a common misunderstanding on the sufficiency of test dimensionality on validity (e.g., combining scores from different content areas). Likewise, the evaluation team provided clarification on test dimensionality in the context of equating methodology.

How reasonable is it to provide a single Science score?

The reporting of a single score for PSU Science was implemented as part of the 2005 admission process. The reporting of a single score responded to a piece of policy mandate from the Board of University Presidents of Chilean Universities (CRUCH). The PSU test battery comprises four standardized instruments: two obligatory tests (Mathematics and Language and Communication) and two optional tests Science and History and Social Sciences). One of the two optional tests must be included as a selection factor for each university career. The Science test has two modules: (1) a common and (2) an elective. Applicants take the common module and must take one of the three elective modules (Biology, Physics and Chemistry). In all the tests forming part of the PSU battery, questions are multiple-choice with five alternative answers, only one of which is correct. The total raw scores of each test are corrected, discounting from the number of correct answers one-fourth of a point for each incorrect one.

The evaluation team considered untenable the PSU practice of producing a single Science score. The fundamental assumption is that this Science score (resulting from adding corrected scores from the scaling of the unique modules to those of the common modules) measures essentially the same construct as the Biology, Physics, and Chemistry tests, with very similar levels of reliability and conditional standard error of measurement. This assumption does not appear to be tenable in light of the nature of the curriculum content covered by each subject areas.

The PSU Science test is based on curriculum frameworks derived from Chile's national high school curriculum. In an attempt to provide a single Science score, Biology scores and Chemistry scores, for example, are linked to Physics scores. The linking is a step to account for differences in the elective modules students happen to choose. The common module is 54 items long and equally taps content curriculum on

the Biology, Physics, and Chemistry taught during the first two years of high school. The elective modules for Biology, Physics, and Chemistry are each 26 items long, and they tap curriculum content taught during the last two years of high school.

Because of this, the process to render a single score for PSU Science tests relies on score conversions that may be subject to misinterpretation due to tests that are not closely parallel. The boundaries of score interpretation and the limitation of the use and comparisons among the modules are neither described nor reported and they should be.

What alternatives are there to replace the current practice of reporting a single score?

The evaluation team considers two competing alternatives to replace the current practice of reporting a single score for the PSU Science test. One alternative considers retaining the current structure of the PSU Science test while targeting the reporting of separate Science scores. Because all college applicants opting for Science take the common module (54 items) and an elective module (26 items), an applicant interested in Biology can receive a content area score (including only Biology items) by considering items from the Biology portion on the common module and the Biology elective portion. The strength we see in this approach is that each of the Science scores can be equated separately and reported on their respective scales without making untenable assumptions on their constructs. Maintenance of the scales across years can be performed with the well-documented equating methodology available to practitioners from the literature (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004).

We also see collateral advantages in the alternative approach. It can be implemented with minimal disruption of the current test development and administration practices. In addition, the same applicant can receive sub-scores on Physics and Chemistry computed by responses to Science items in the common module. While the content area score can be considered part of admission criteria for Biology-oriented careers, sub-scores can be valuable aids to provide formative information to test takers to be used to improve Science achievement outside the context of university admissions testing, e.g. incoming university students entry level knowledge in Science.

One limitation concerning the validity of Science test scores is due to decisions that were made when the domain and intended population of the PSU Science tests were initially defined. From the beginning, the Science test was informed by the national high school Science curriculum in two ways: the common module of Science would measure content taught in all schools during the first two years of high school; the elective module of Science would measure content taught in the Scientific-Humanistic curricular branch after the second year of high school. As increasing numbers of Technical-Professional students begin to take the PSU Science test, it raises doubts on meaning and fairness of the total Science score for this group.

The other alternative considers abandoning current PSU Science test design to allow for separate standalone Science tests for Biology, Physics and Chemistry. Under this alternative PSU Science assessment frameworks should be develop to account for characteristics of the population of test takers, the purpose of the test, and the intended use of the test scores. In addition, maintenance of the scales across years can be performed with the equating methodology well documented in the literature (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). These two advantages

are major improvements on the current issues that the PSU Science test faces. The downside of the option is logistic. This alternative would require a complete overhaul of the PSU Science tests and related assessment policy (e.g., use of a single score for PSU Science composed of common and elective portions).

Dimensionality

Dimensionality is a relevant aspect on psychological testing, and it has significance for validity in educational testing when exploring the homogeneity of a test domain. Educational testing derives meaning from test scores from test frameworks (typically curriculum and test items) (AERA, APA, & NCME, 1999; Standards 3.6, 3.11). In this context, test alignment to curriculum provides the basis for interpreting the meaning of test scores, and dimensionality analyses provide information on technical adequacy of the test (e.g., equating).

In the context of the PSU, the domain of the Science test is based on Chile's high school curriculum frameworks, and a single Science score is attained with an observe score regression-type approach. As part of the Objective 2.2 of this evaluation, the evaluation team documented the presence of low levels of alignment of the PSU Science test. In addition, as part of the Objective 2.1 of this evaluation, the evaluation team documented evidence of a single factor for each of the Science content areas and the Science common set.

Dimensionality analyses are also valuable to assess validity of IRT local independence assumption. However, in Chile the PSU test, including other PSU tests, is not equated. Instead, the PSU Science test uses an observe score regression-type approach to combine biology, chemistry, and physics score into a single score. Although the statistical assumption of the computations does not depend on dimensionality, validity of the single score does.

The evaluation team considers the reporting of a single PSU Science score to be untenable because it relies on a questionable assumption of equivalence (e.g., meaning) of the part-test scores (Biology, Physics and Chemistry).

GENERAL DESCRIPTION (PART II)

The authors of the process for "equating" PSU Science test scores claimed that the process is a variant of the nonlinear methods that establish score equivalence of the elective modules by conditioning their observed scores (elective tests) on scores from the common module. The process intends to compensate for differences on the elective modules difficulty and groups' characteristic (applicants choose to take one of the tests among the elective modules). The process comprises three stages: (1) regression of elective scores on common scores, (2) interpolation of scores, and (3) adjustment of elective scores. Appendix A shows equations for the steps of the process.

The sign-off process involves independent replications between DEMRE (*unidad de estudios de investigación*) and PSU Technical Advisory Committee (CTA). The process requires comparisons of chosen nodal points and linked scores. The former comparison has been the one yielding occasional discrepancies which, once resolved, result in matched results and process sign-off. The discrepancy resolution was described to be a participative process in which DEMRE staff and CTA discussed their results and agreed upon the subsequent course of action. Once the discrepancies have been reconciled, the standard process is followed to render linking outcomes. The sign-off process currently is not documented as part of DEMRE quality assurance processes.

The single score of PSU Science test is the sum of scores (after correcting for guessing) attained from the common module and the adjusted elective score. The single Science score is normalized using a standardized normal curve. PSU scale scores are derived through linear transformations of the normalized scores to render a scale with a mean of 500 and a standard deviation of 110. A small correction is performed on the scores at the distribution extremes, with the purpose of establishing a minimum of 150 points and a maximum of 850 points on the PSU scale. The above two steps are common processes followed for reporting PSU tests scores.

Population of test takers

The larger influx of PSU test takers took place between the 2006 and 2007 admissions processes. As a result of national educational policy, the government financed the registration fee to the PSU for all students graduating from municipal and subsidized high schools. Historically, for the Science test, the elective module most administered has been Biology, but the one registering the greatest percentage applicant increase has been Chemistry. (See Table 91.)

Table 91: Students taking PSU Science Tests across Admission Years

	2004	2005	2006	2007	2008	2009
Science	84,214	94,139	99,530	113,530	120,657	138,572
Elective: Biology	49,016	54,451	55,740	62,386	64,756	76,555
Elective: Physics	19,827	20,722	21,248	23,264	24,341	27,608
Elective: Chemistry	15,371	18,966	22,542	27,880	31,560	34,409

Characteristics of Test Score Distributions

Table 92 through Table 94 show distributions of PSU Science test scores by PSU Science tests. The tables provide descriptive statistics computed for tests and applicants of the 2010 admission process (DEMRE, 2010). The statistical summaries shown are for the tests comprising the common and elective portions. Three aspects of the table are of particular importance. First, PSU Biology (n=81,301) captures the larger number of applicants among the three Science tests, followed by Chemistry (n=31,799) and Physics (n=28,900). Second, the Chemistry test shows the highest average revised score (i.e., raw score corrected for guessing) (M=23.32; SD=18.70), followed by Physics (M=20.66; SD=18.50), and Biology (M=15.85; SD=16.60). Third, Chemistry shows the largest standard measurement error (SEM= 4.27), followed by Physics (SEM=4.25), and Biology (SEM=3.75).

There is evidence of high degrees of linear relationship between Science scores from the common portion and the elective portion. In 2010, Pearson *r*-correlations ranged from 0.867 (common-elective Physics) to 0.870 (common-elective Biology).

Table 92: Descriptive Summary for 2010 Admissions Process by PSU Science Test – Biology

Total applicants	N	81,301
Total questions	k	80
Average revised score	μ	15.85
Standard deviation of revised score	σ	16.60
Average degree of difficulty	%	30
Average index of homogeneity	rb	0.615
Reliability (Alpha Cronbach)	rtt	0.95
Measurement error	E.M.	3.75
Asymmetry Coefficient	As	1.34
Kurtosis Coefficient	K	1.40

Table 93: Descriptive Summary for 2010 Admissions Process by PSU Science Test – Physics

Total applicants	N	28,900
Total questions	k	80
Average revised score	μ	20.66
Standard deviation of revised score	σ	18.50
Average degree of difficulty	%	29
Average index of homogeneity	rb	0.631
Reliability (Alpha Cronbach)	rtt	0.95
Measurement error	E.M.	4.25
Asymmetry Coefficient	As	1.03
Kurtosis Coefficient	K	0.27

Table 94: Descriptive Summary for 2010 Admissions Process by PSU Science Test – Chemistry

Total applicants	N	31,799
Total questions	k	80
Average revised score	μ	23.32
Standard deviation of revised score	σ	18.70
Average degree of difficulty	%	30
Average index of homogeneity	rb	0.613
Reliability (Alpha Cronbach)	rtt	0.95
Measurement error	E.M.	4.27
Asymmetry Coefficient	As	0.83
Kurtosis Coefficient	K	-0.10

EVALUATION

The process that is being followed to render a single Science score is NOT equating in the strict sense defined by Kolen and Brennan (2004) because the examinees take tests with different content based on the optional sections (alternative modules). Because these alternative modules bring content differences, scores for students taking the different optional modules cannot be considered to be equated. However, such scores can be referred to as "linked," and the process followed referred to as "linking." Terminology associated with the single score in Science needs to be changed from "equating" to "linking."

The process currently used to develop a score for each examinee involves linking the score on each optional section to the score on the common portion using a nonlinear regression method. The single score is the sum of the score on the common portion and the linked score on the optional section. Separate nonlinear regressions are used for the three optional sections to derive the linked score portion of the single score. This process is described in detail in technical documentation.

The nonlinear regression procedure uses a fixed set of nodes (scores on the common portion) of the tests and finds a nonlinear regression of optional scores for the common section scores. The nonlinear regression procedure appears to fit a linear relationship between nodes and appears to result in a piecewise linear function.

No statistical rationale is given for the particular choice of nodes in the regression procedure. In addition, no statistical rationale is given for the use of what is apparently a piecewise linear regression function. Alternate sets of nodes would likely lead to different linking results. In addition, a cubic spline regression procedure (i.e., smoothing splines) likely would be an improvement to the procedure used here because cubic splines (see Kolen and Brennan, 2004, for a discussion of the use of cubic splines in equating) produce a continuous curvilinear regression function and criteria exist in the literature for choosing nodes and smoothing parameters.

No statistical rationale is provided for calculating a single score by summing the scores on the common section and scores on the linked optional section. The correlation between the common portion and the optional portion will have a substantial effect on the variability of total scores. For these tests, the correlations between the common portion and the optional portions were nearly equal for the three optional sections, so the variability of total scores likely were similar. However, if at some point in time these correlations were to differ substantially, the variability of the summed scores could be quite different for examinees taking different optional sections.

The statistical criteria used are not stated for the procedure, which gives rise to the following questions: What is the method intended to accomplish from a statistical or psychometric perspective? What statistical or psychometric assumptions are being made?

From the analyses presented, it is difficult to ascertain the extent of comparability of total scores for students taking different modules. Based on the way the procedure is implemented, it appears that the single score for a student who took a Biology optional module is considered to be comparable to a student who took a Chemistry or Physics optional module. The rationale for this comparability is not clear.

It would be informative to regress outcome variables such as college grade point average (on comparable college science courses) on the Science single score for the groups taking the Biology, Chemistry and Physics optional modules. If the total scores are comparable, these three regressions should result on regression coefficients approximately equal.

Table 95 shows a summary evaluation for the PSU Science equating process. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing aspects for fine grain improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 95: PSU Science Equating Process

Facet 1: PSU Science Equating: Process	
1. Describe PSU process to derive a single score for Science test.	
2. Follow up questions (if needed and applicable) What is the Psychometric Foundation of the process?	Rating
• Is there a pre-established process to render a single score for Science?	C (4.11)
• What is the statistical criterion to meet with the process?	C (4.11)
• What are the characteristics of the score distributions targeted by the process?	C (4.10)
• What type of statistical assumptions the process made?	D*
• Who participates in the equating?	E
• What data collection design is followed?	E

• Is PSU Science test IRT dimensionality relevant to the process? Why yes or why no?	C (4.17)
• What is the sign-off process, and who participates in it?	E

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

EVALUATION

Coefficient alpha is used as the reliability coefficient in documentation of PSU Science. Strictly speaking, this coefficient is appropriate only when the total score is the sum of the item scores. For the PSU, the common item scores are added to the linked optional section scores which is NOT the sum of the item scores. Although coefficient alpha might be a reasonable estimate of reliability, it is not clear that it is appropriate. A more appropriate procedure would be to use reliability estimation methods for composite scores.

It is unclear how the overall standard error of measurement is being calculated, though if it makes use of coefficient alpha then it is not clear that it is appropriate. Conditional standard errors of measurement were not provided in the information that we received; these should be provided to test users.

In any case, separate reliability coefficients and standard errors of measurement should be reported for examinees taking each of the three optional modules.

Table 96 shows a summary evaluation for reliability and conditional standard error of measurement of PSU Science score. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing aspects for fine grain improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 96: PSU Science Equating Process (Reliability and CSEM)

Facet 2: PSU Science Equating: Reliability and Conditional Standard Error	
1. Describe PSU process to derive reliability and conditional standard error of the single score for Science test.	
2. Follow up questions (if needed and applicable)	Rating
• Is the pre-established process to estimate reliability of a single score for Science?	C (4.11)
• What is the process to estimate reliability of Science single score? • Is there a pre-established criterion to evaluate and interpret magnitude of reliability coefficients?	C (4.11)
• Is the pre-established process to estimate conditional standard error of measurement (CSEM) for Science single score?	B (4.10, 4.11)
• What is the process to estimate CSEM for the Science single score? • Is there a pre-established criterion to evaluate and interpret magnitude of the conditional standard error of measurement for Science single scores?	B (4.10, 4.11)

EVALUATION

Equating error is an important component when evaluating the accuracy of equating functions. International standards (AERA, APA, & NCME, 1999) endorse the need to report estimates of sampling error present in scaling functions. In theory, scaling should provide accurate functions for any sample of examinees from an intended population of examinees. In the context of the PSU, Science tests are scaled to achieve comparability. This type of scaling is followed to link measures of academic achievement in several content areas of Chile's national high school curriculum (e.g., Physics, Chemistry, and Biology). Relative to equating, scaling to achieve comparability renders score conversion tables that are weaker, though they need to be technically sound to satisfy comparability goals. Scaling to achieve comparability results cannot be assumed to be generalized across time, subpopulation of test takers, and tests.

No information documenting linking ("equating") error is provided. For the method used here, it would be advisable to use a bootstrap method (see Kolen and Brennan,

2004, for a description of using the bootstrap in equating) to estimate such standard errors.

The evaluation team considers the current status of PSU Science equating showing lack of reported accuracy of scores a condition in need for improvement. Without this and other improvements that have been raised for other facets, the evaluation team declares unacceptable the current status of PSU Science single score.

Table 97 shows a summary evaluation for equating error of PSU Science. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing aspects for fine grain improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 97: PSU Science Equating (Equating Error)

Facet 3: PSU Science Equating: Equating error	
1. Describe PSU process to estimate equating error for the process.	
Follow up questions (if needed and applicable)	Rating
• Is the pre-established process allows estimating equating/linking error?	C (4.10)
• What is the process to estimate equating/linking error for Science single score? • Is there a pre-established criterion to evaluate and interpret magnitude of the equating/linking error entered as part of Science equating?	C (4.10)

EVALUATION

A variety of procedures exist that could be used to conduct linking of scores for students taking different optional sections. The method that is currently used links scores from each optional section to the score on the common portion. These linked scores are added to the scores on the common portion. As mentioned earlier, this sort of process could lead to scores being unequally variable if the correlation between common item scores and optional section scores differ. Interestingly, data at hand showed comparable degrees of intercorrelation but such a check should be incorporated into the current approach to link PSU Science test scores.

An alternative would be to calculate a total score by summing the raw scores on the common portion and the optional section separately for the group of examinees administered each optional section. These total scores could be linked to one another using the scores on the common portion as an internal set of common items (Kolen & Brennan, 2004). Chained equipercentile, frequency estimation equipercentile, chained linear, or Tucker linear methods (all described in Kolen & Brennan, 2004) could be used to link total scores for each group to scores on a base test (e.g., the common portion plus Biology). These standard methods are described in Kolen and Brennan (2004), are fully documented, and have been used in a variety of testing programs. The benefit of using these standard procedures is that their assumptions have been carefully and comprehensively described in the literature. If equipercentile methods were used, then a smoothing procedure (Kolen & Brennan, 2004) could be used to reduce sampling error. In addition, bootstrap procedures could be used to estimate standard errors of the linked scores.

There is substantial evidence in the literature reviewed by Kolen and Brennan (2004) that when groups differ substantially in ability, the linking results are highly dependent on the group used to conduct the linking. With the PSU, based on scores on the common portion, it appears that the group of students taking the Biology optional modules is substantially less able than those taking the Chemistry or Physics optional modules. Due to these substantial differences, the method used to link the scores likely will have a substantial effect on the scores that result from the linking.

With any of these methods there is an implicit assumption that total scores are in some sense equivalent, regardless of the optional module taken. This assumption, which appears to be unrealistic, needs to be thoroughly investigated. For any procedures, including these standard ones, it is important to check on score comparability, even if the more standard procedures are used. As suggested earlier,

it would be informative to regress outcome variables such as college grade point average (on comparable Science courses) on the total score separately for the groups taking the Biology, Chemistry, and Physics optional modules using whatever method is chosen. If the total scores are comparable, these three regressions should be approximately equal.

Table 98 shows a summary evaluation for the model used to render a single score for PSU Science. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing aspects for fine grain improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 98: PSU Science "Equating" (Models)

Facet 4: PSU Science Equating: Models	
1. Describe rationality behind use of models for PSU Science equating process.	
2. Follow up questions (if needed and applicable)	Rating
• Is the PSU Science equating process compensatory/non compensatory in nature?	C (4.10)
• What is the rationale to choose a non-linear model?	E
• What is the rationale to choose nodes instead of relying on span of test scores for the PSU Science common portion?	C (4.10)
• What are the effects of choosing a given PSU Science test instead of another as a base for one year?	C (4.10)
• Has the model thoroughly researched and documented outside of the PSU test context?	C (4.10)
• Has the model been compared to competing models available from literature and used internationally involving non-equivalent groups and observe scores (e.g., Tucker, Levine, frequency estimation equipercentile equating)?	D*

*****A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

EVALUATION

Sound testing practices require continuous quality improvements and repeated reviews of decisions made during the course of the operational program. There is little evidence that processes to render a single score for Science tests have undergone thorough revisions and checks to improve decisions made early in 2005. The evaluation team considers this limitation to be doable and encourages stakeholders to revisit their processes in light of the recommendations from this evaluation.

Table 99 shows a summary evaluation for process maintenance and improvement for PSU Science. The purpose of the table is to provide an analytic evaluation to fine-tune the above holistic evaluation with purposes of pinpointing aspects for fine grain improvement decisions.

The analytical evaluation reported in this section makes use of the following coding schema. In addition, the table shows, under column labeled "Rating," the list of professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 99: PSU Science Equating (Process maintenance and improvement)

Facet 5: PSU Science Equating: Process maintenance and improvement	
1. Describe process maintenance and process improvement plans.	
2. Follow up questions (if needed and applicable)	Rating
• Are the specifications developed surrounding the process to evaluate and detect improvement needs for PSU Science equating?	C (4.17)
• Are there plans and schedules set to perform process evaluation and improvement needs?	C (4.17)
• Are there logistics put in place to close gaps between actual and desirable states of improvement?	C (4.17)

RECOMMENDATIONS

1. The international PSU evaluation team considers the rationale for PSU Science score linking to fall below international standards. The process is not only incorrectly labeled but also the documentation is incomplete and the evidence of technical adequacy insufficient for high-stakes decisions. The evaluation team recommends developing separate Science tests for Biology, Chemistry and Physics with specific purposes and intended populations in mind so that the scores would have unambiguous meaning. Each of these tests would be reported on separate PSU scales, following standard processes already available for PSU tests. Furthermore, once the current cumbersome linking process had been replaced, the year-to-year maintenance of the new PSU Science scales through equating would be more rigorous and, hence, more defensible.
2. Until our recommendation can be implemented, there is a need for more documentation for the current Science tests that informs the public and technical reviewers alike about the current policy decision to report a single score for Science, its rationale, and the research that informed that decision. The practice of linking test scores from different content areas has been performed to achieve comparability through scaling. This form of linking is weak when compared to equating. For that reason, evidence of the generalization of conversion tables should be provided for sub-groups, occasions, and tests. Other recommendations for the current PSU Science tests include the following:
    g. Refer to the process as "linking" rather than "equating."
    h. Link total scores, rather than using the current process of linking scores on optional sections and then summing the linked scores with the scores on the common portion.
    i. Consider using standard linking methods, such as chained equipercentile and frequency estimation equipercentile. Smoothing methods should be used with these procedures. Thoroughly compare the results for all methods considered. Provide statistical and psychometric criteria that indicate what the procedure is intended to accomplish.
    j. With the current linking methods there is an implicit assumption that total scores are in some sense equivalent regardless of the optional module taken. This assumption, which appears to be unrealistic, needs to be thoroughly investigated. For any procedures considered, including these standard ones (e.g., chained equipercentile and frequency estimation equipercentile), it is important to check on score comparability. Regress outcome variables such as college grade point average on total score for the groups taking the Biology, Chemistry, and Physics optional modules. If the total scores are comparable, these three regressions should be approximately equal.
    k. Estimate reliability, standard errors of measurement, and conditional standard errors of measurement using procedures that have been developed for composite scores. Calculate standard errors of the linked scores using bootstrap procedures.
    l. Document processes describing quality assurance and quality checks for PSU Science score linking.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Comité Técnico Asesor. (2010). *Procedimiento de equating para la prueba de ciencias PSU*. Santiago: Universidad de Chile.

DEMRE. (2010). *Prueba de selección universitaria (PSU): Antecedentes y especificaciones técnicas*. Santiago: Universidad de Chile.

DEMRE. (2010). *Pruebas de selección universitaria: Proceso de admisión 2010*. Informe Técnico. Santiago: Universidad de Chile.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Norwell MA: Kluwer Academic Press.

Kolen, M. (2006). Scaling and norming. In R. L. Brennan (Ed.) *Educational measurement* (4th ed. pp. 155-186). Westpoint, CT: American Council on Education and Praeger Publications.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Objective 1.3. Evaluation of IRT models for item calibration, test development and equating purposes

Item response theory (IRT) provides an alternative powerful framework to classical test theory (CTT) for multiple aspects of test construction, item calibration, and equating process. Internationally, assessment programs are moving more toward using the IRT framework for the advantages it provides compared to the CTT framework.

Over the course of this evaluation, the evaluation team has gained deeper understanding on the surrounding PSU psychometric processes (e.g., item calibration and test equating). We have found that in Chile, year-to-year equating of PSU test administrations is not performed to maintain the PSU score scale. In addition, IRT item calibration processes are deficient.

We have found that in Chile, the documentation exists for processes that are not taking place for the PSU (e.g., equating) and for processes that are ill developed (e.g., IRT calibration). These are the two criticisms that we emphasize, and they need to be considered when reading this section of our evaluation.

To comply with the evaluation, we captured from the interviews and documentation information and presented its analysis as part of this report. For this effort, we involved international standards.

Nevertheless, the fact remains that in Chile, equating does not take place for the PSU and the IRT calibration that does occur is deficient.

The evaluation team met with relevant stakeholders from DEMRE on March 26, 2012, to participate in a demonstration of IRT calibration with a synthetic data set. The purpose of the demonstration was to gain deeper understanding of DEMRE's approach to IRT item calibration and participants' decisions and rationale for utilizing the models and processes. The meeting was geared toward covering the following aspects:

- Gain details on ways that DEMRE implements IRT item calibration
    - Control card
    - IRT model
    - Data fit
    - Data collection
    - Quality control
- Perform IRT item calibration with synthetic data set
- Year-to-year PSU equating (Not applicable)
- Role of anchor set on year-to-year equating (Not applicable)
- Process maintenance and evaluation

The evaluation team read DEMRE documentation and related information (i.e., control card) about PSU calibration process. In preparation for the meeting, the evaluation team created a synthetic data array with 3000 simulees and 50 items with known item parameter properties utilizing standard simulation methodology.

The meeting was performed within DEMRE offices following an agreed-upon schedule for the visit. Outcomes of the IRT item calibration process were collected from the meeting for later comparison with the intended item parameters underlying the synthetic data set.

The following DEMRE staff participated in the interviews:

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees with process that the evaluation team followed in the working session.

The following subsection contains the results of the evaluation for Objective 1.3.

The international evaluation team relied on professional standards during their appraisal of the merit and worth of the PSU process involving IRT models for item calibration, test development, and equating purposes. A framework for evaluating PSU approaches for reliability and standard error of measurement has been developed from the following set of standards found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Standard 3.9

> When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. […] When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented. (p. 45)

Standard 4.10

> A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably. In some cases, direct evidence of scores equivalence may be provided. In other cases, evidence may come from a demonstration that the theoretical assumptions underlying procedures for establishing score comparability have been sufficiently satisfied. The specific rationale and the evidence required will depend in part on the intended uses for which score equivalence is claimed. (p. 57)

Standard 4.11

> When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions. (p. 57)

Standard 4.13

> In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the form being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-

based and classical equating studies, the representativeness and psychometric characteristics of anchor items should be presented. (p. 58)

GENERAL DESCRIPTION

University admissions testing is a high-stakes endeavor with highly developed systems to protect the integrity of test items and test scores. In Chile, it has been judged critical not to embed field test items into the operational administration of the PSU battery. Instead, separate independent "standalone" administrations of field test forms have been conducted to collect performance statistics on field test items for the purpose of item banking (DEMRE, 2011a). DEMRE has put some documentation describing the sampling plan for field testing items. The main highlights on the sampling plan are: (1) the target population of field test participants comprises all students graduating from high school, (2) stratification variables includes region, high school curriculum modality, type of high school, and gender, (3) multiple field test forms, and (4) sample sizes per field test form that target 1500 students for non-Science tests and about 1300 students for Science tests. (Note: For additional details, please consult Objective 1.1.b. of this evaluation report.)

DEMRE has documented the process for selecting anchor items for field testing activities (DEMRE 2011b). This documentation states that the goal of the anchor items is to equate pretest measures with operational administrations of the PSU from previous years. In this respect, DEMRE states:

> [S]elección de ítemes de anclaje para efectuar el proceso de equiparación entre el pretest que se desea analizar y la prueba oficial realizada dos años previos al proceso de ensamblaje. [...] criterios están basados en los índices de discriminación y de dificultad de los ítemes, desde la perspectiva de la Teoría Clásica de Test (TCT). (DEMRE, 2011b. p. 3)

But, in reality, test equating is not performed. In this regard, the anchor items are essentially useless since they do not fulfill their purpose.

Our analysis of the processes outlined in the documentation of 'anchor items' makes us to pinpoint several deficiencies. Table 100 shows a summary of the documented criteria for selecting anchor items for field test purposes (DEMRE, 2011c). The documentation depicts anchor item sets to be short and unrepresentative of the full test. In each of the PSU tests analyzed, the descriptions of the anchor set fell short of international standards for anchor length and representativeness (Kolen & Brennan, 2004). Furthermore, the documented statistical characteristics of the anchor sets (item difficulty and item discrimination) are insufficient to be evaluated. Interestingly, the documentation uses classical test theory item difficulty and item discrimination indices to define intended characteristics of anchor sets for an IRT context without referencing target levels of test difficulty and precision. In addition, there is a lack of information about the actions taken, at least from a design perspective, to reduce context effects on the performance of anchor items and to institute processes that check for anchor item parameter drift.

Table 100: Anchor Set Characteristics by PSU Test

PSU Test	Item Difficulty*	e	Test length	Anchor length
Language and communication	Spread between (≤15.0 and ≥ 70)	At least 0.50	80	8
Mathematics	Spread between (≤15.0 and ≥ 70)	At least 0.50	70	7
History and Social Sciences	Spread between (≤15.0 and ≥ 70)	At least 0.50	75	7
Science (Common)	Spread between (≤15.0 and ≥ 70)	At least 0.50	54	6
Science (Biology)	Spread between (≤15.0 and ≥ 70)	At least 0.50	26	3
Science (Physics)	Spread between (≤15.0 and ≥ 70)	At least 0.50	26	3
Science (Chemistry)	Spread between (≤15.0 and ≥ 70)	At least 0.50	26	3

(*) Percent correct
(**) Biserial correlation

The PSU field test item response theory (IRT) calibration is performed with the 2-parameter-logistic (2PL) model with BILOG 3.11 (DEMRE 2011c). The 2PL model defines the probability of correctly responding to an item as a function of the test taker's ability, item difficulty, and item discrimination parameters. This model assumes that guessing is of no concern for understanding item performance and modeling test taker's ability. In other words, under the 2PL model the probability of getting an item correct by a very low ability test taker is zero.

In order to replenish the PSU item bank as new tests are developed each year, newly created items are field tested and calibrated. DEMRE uses BILOG 3.11, a standard, commercially available, IRT software, to perform the PSU field test calibration. Once the field test items are administered, items are calibrated to estimate the following psychometric features: (1) item characteristic curve, (2) item difficulty, (3) item discrimination, and (4) item information function (DEMRE, 2001a). Calibrations are performed on a yearly basis and scaled on the default scale centered at 0 with a standard deviation of one. That is, for each PSU test a scored field test data file (1=correct, 0=wrong, 9=Omit, " "= Blank) is processed with BILOG 3.11 and resulting item parameters estimates are loaded into the PSU item bank.

The following figure shows a schematic representation of commands and key words present in the BILOG 3.11 control card used for PSU field test item calibrations. Commands and key words in the figure were chosen with the evaluation goal in mind.

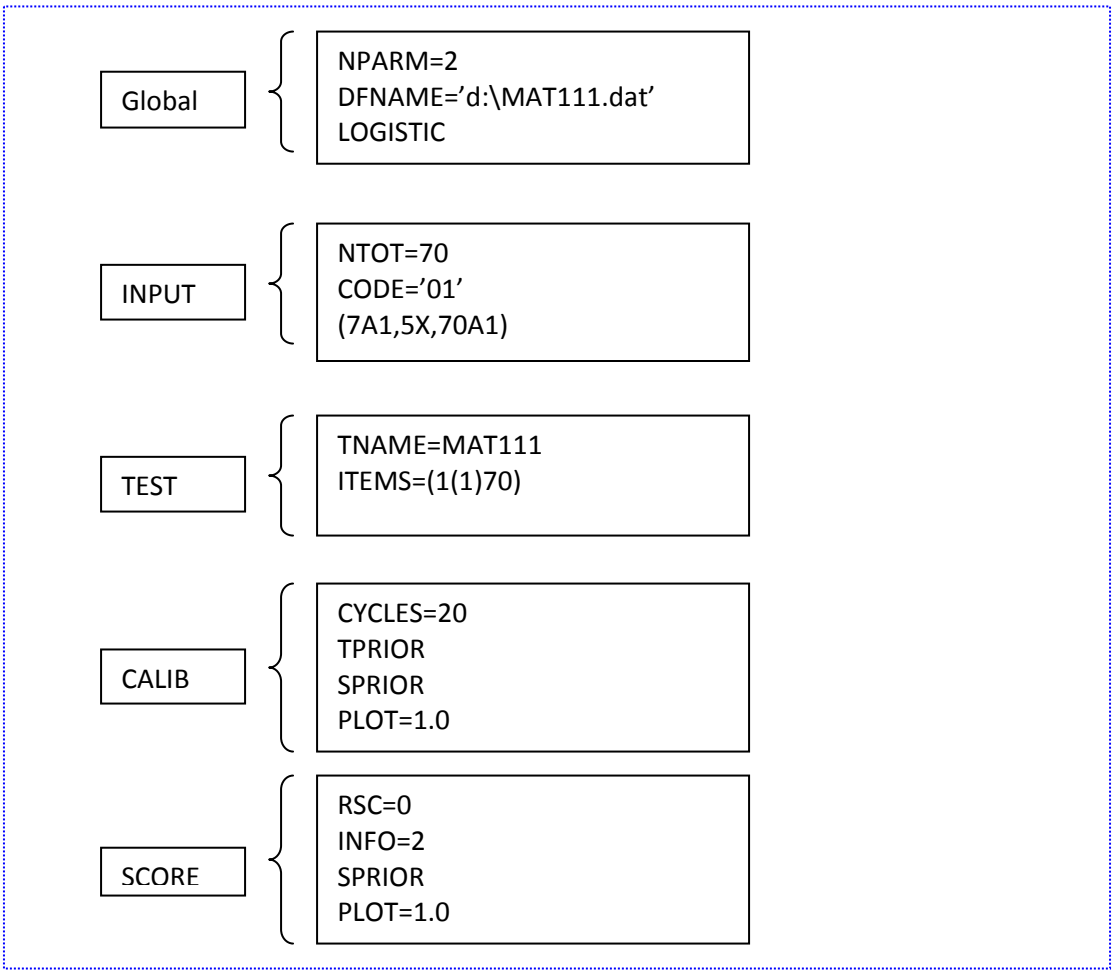


Figure 36: Selected BILOG 3.11 Commands and Key Words on PSU Mathematics Control Card

EVALUATION

The evaluation team after considering all facets and elements inspected for Objective 1.3 rejects the IRT documentation and processes currently used in the PSU testing program. The documentation reviewed is misleading. The processes taking place need to be labeled properly, and processes that are not taking place need to be identified. For example, the evaluation team became aware of the misleading use of equating terminology (e.g. anchor set, model fit) and corrected their initial understanding. Specifically, the PSU testing program does not maintain a reporting scale on a yearly basis nor does it calibrate field test items to a common scale.

A clear description of a process to calibrate field test items is an important practice often found in mature testing programs involving high-stakes assessments. When developing and assessing test items using item response theory (IRT), practitioners design processes in a way that field test item parameters can share a common scale with existing field test item parameters in item banks. Producing field tested item parameter estimates within a common scale allows for an "apples-to-apples" comparisons of item parameters and related

IRT information, both among multiple field test forms in any given year and across years of operational administrations.

In the context of Chile's university admissions process, the evaluation team recommends correcting inaccuracies present in the PSU documentation for field test equating. The existing process for developing an anchor set, as described in official documents, is unconventional, and it does not reflect international standards. For example, the length of the anchor set is below the standard ratio between anchors set and test length. The ratio mentioned in the documentation for the PSU does not reach the minimum of 25 percent of the total test length (Kolen & Brennan, 2004). In addition, the anchor sets are constructed without guidelines on how to achieve content representation. As a result, these anchor sets fall short of completely and accurately representing total test characteristics.

Within the framework of international standards, practitioners bolster the psychometric documentation around their calibration efforts with sufficient information that allows, for example, (1) staff from the testing program to perform approved processes for scaling, and (2) external agencies to perform an evaluation of processes. Typically, the psychometric documentation is developed and maintained throughout the duration of the testing program.

There are several sections often present in the psychometric documentation that cover the breadth and depth of psychometric processes. Critical aspects on the documentation are: (1) descriptions of major modifications on the assessment, (2) a description of data structure congruent to data collection process (e.g., incomplete data matrix and matrix sampling design), (3) a description of psychometric models and rationale for their use, (4) approaches to estimate scale transformation functions, (5) processes to study stability of anchor item sets, (6) processes to evaluate accuracy of results, and (7) quality control processes with corresponding schedules.

There are several ways that the PSU testing program does not attain a level of analysis generally expected when equating a high-stakes assessment. The most striking irregularity was the misleading information found in the documentation of the assessment program: specifically, the information concerning year-to-year equating and field test calibrations. The international evaluation team reiterates that these activities are not taking place, even though the documentation provided seems to indicate that they are. The evaluation team emphasizes the necessity for equating activities to be carried-out for the PSU tests.

The PSU testing program also does not check the fundamental assumption of the IRT model in a coherent and clear manner. In an ideal world, correction for guessing removes guessing from applicants' responses, and thus a 2PL formulation becomes suitable. In the real world, a 3PL formulation brings an extra-parameter (the *c*-parameter) to account for guessing tendencies without a need for involving the controversial correction for guessing formula. In addition, the model assumption behind the 2PLM is the presence of a strong dimension explaining covariance among test items. Technically sound testing programs routinely check the unidimensionality assumption by utilizing approaches such as item level factor analyses. Once that dimensionality check has been cleared, testing programs will then screen for model fit to data. In the context of IRT, Chi-square index (likelihood ratio) is often available as part of the psychometric software and used to judged model fit. In addition, graphical analyses of expected and observed probability of correctly answering an item are often inspected as part of data fit analyses. The latter approach has the advantage of being intuitive and easy to explain. The former approach has the disadvantage of becoming too powerful with large sample sizes. Nevertheless, the international evaluation team believes the irregularities noted above could be redressed if external review audits are developed and reinforced.

Any external review audit would ask the practitioners to provide a detailed narrative of test construction specifications for field test purposes. Test construction specifications provide information outlining characteristics of the field test (or operational test), which is important information for the staff participating in (or reviewing) the calibration effort. Important sections of these documents are directed toward specifying psychometric targets for the test. Of particular interest are the descriptions of content and the statistical representation of anchor item set, the sequencing of anchor items in the form, and the status of anchor item sets counting or not counting toward students' total scores. When multiple field test forms are involved, the availability of such a document becomes a critical aspect of a well-established testing program.

During the review of the documentation and the interviews, the evaluation team became aware of the loose usage of equating terminology and processes. In the context of equating the PSU testing program, there is a discrepancy between what is documented and what is practiced. From the interviews, it became evident what the purpose of the so-called anchor set has been in the program. For example, the term "anchor set" is being used to refer to a group of items that are added to a field test form with a purpose other than the calibration of field test administrations to a common scale (AERA, APA, & NCME, 1999. Standard 4.13).

Table 101 shows a summary of the evaluation of the IRT methods for calibration of the PSU item responses and the IRT approaches for PSU test core equating. Under the column labeled "Rating" found in the table, the evaluation team lists parenthetically the particular professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As

well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 101: Summary Evaluation of IRT Calibration and Equating (FACET 1)

Facet 1: PSU Item calibration IRT model	
1. Describe processes for PSU Field Test IRT Calibrations	
2. Follow up questions (if needed and if applicable)	Rating
a. What IRT model(s) are used to calibrate PSU field test items? • I am hearing IRT 2 PLM is utilized to calibrate PSU field test items. What data collection design is followed (e.g., common anchor set)? • What variables are included in the data file? • What type of item level data is present in the data file? • What consideration(s) were given to alternative IRT models (e.g., 1PL and 3PLM) when selecting the IRT calibration approach? • What consideration(s) were given to unidimensionality assumption?	C (4.13)
b. What types of outcomes are sought for the IRT calibration of PSU field test items (e.g., ICC and IIF)?	C (4.11)
c. What are the characteristics of scale involved in the IRT calibration of PSU field test items? ▪ Does control card reflect PSU scoring (e.g., correction for guessing)? ▪ Does control card show evidence of item parameters to be calibrated to a pre-existing scale?	B (3.9, 4.13)

EVALUATION

In Chile, the documentation of IRT approaches to PSU item calibration and PSU test equating is emerging in several ways that could show improvement over the existing practices, which are insufficient. To date, PSU item calibration analyses of BILOG 3.11 control cards showed important drawbacks affecting the validity of estimated item parameters. There is no linking mechanism built into the calibration process of FT items. Neither the documentation surrounding anchor set selection nor the actual control cards used in calibration efforts indicated that field test items are calibrated to a common scale, which would enable comparisons of item parameters across forms and across years. This is an important limitation that hinders not only the comparability of item parameters available in the PSU item bank, but also the test construction efforts informed with IRT parameters.

PSU item calibration is performed under the 2PL model under the assumption that the correction for guessing has removed the guessing component out of applicants' item responses. The evaluation team found several areas for improvement when analyzing DEMRE process to calibrate PSU items with 2PL model. In Chile, university admissions tests are corrected for guessing and the correction applies for wrong answers but not for omitted responses. Test directions state the use of correction for guessing. PSU item calibration is performed with 2PL IRT and valid data for correct and incorrect item responses. The calibration process leaves out scores of "9" (defining omitted responses) from the calibration process. Figure 36 shows that under the INPUT command, valid scores are 0 and 1. The decision to focus just on correct/incorrect item responses might result on a set of estimated item parameters that is dependent on an analyst's decision to ignore the occurrence of omitted responses. In addition, there is no evidence surrounding superiority

of the 2PL model over the 3PL model in the context of PSU item calibrations. The rationale behind the control of guessing by the correction scoring gets lost when decisions are made to calibrate data sets with correct and incorrect responses as the valid scores. Thus, the evaluation team strongly recommends validating the adequacy of the 2PL for modeling PSU items. Similarly, the evaluation team recommends developing documentation of item calibrations that are available to staff who participate in the calibration of items. Such documentations could be used to train staff on PSU item calibration processes that have been approved by DEMRE and the technical advisory committee (CTA).

There is major drawback in the current PSU item calibration process. From PSU item calibration control card currently used, year-to-year calibrations are not referenced to a common scale. The absence of such a calibration effort creates concerns about the comparability of item parameters, the associated test construction activities, and the quality of item bank information. Dimensionality analyses are also absent from PSU item calibration process. It would be commendable to add this type of analysis to check model assumptions.

Table 102 shows a summary evaluation of the PSU IRT model fit and data collection design. Part of the PSU item calibration process includes the generation of item fit indices, item characteristic curves, and item information functions for PSU item calibration. These pieces of information are easily accessible from the calibration software and typically scrutinized by DEMRE analysts. The evaluation team encourages a continuation of these practices and a clear documentation of the process instituted to maintain records of misfitting items. It would be also commendable to add model fit check and dimensionality analyses. It would be commendable to add documentation to the existing processes, particularly on the areas of model fit and model assumption check.

Data collection design is a key component on field testing efforts. Objective 1.1.b. provided an analysis of the sampling process used to collect data for field test analyses. The sampling plan should provide a rationale behind relevant variables. The sampling plan can be improved by adding sections to describe the process check-lists intended to evaluate implementation of the FT sampling plan. The sampling plan may also need to be revised to accommodate a shift in the definition of relevant sampling variables. For example, parents' education and family income can be used to derive an indicator of socio-economic status.

Table 102: Summary Evaluation of IRT Calibration and Equating (FACET 2)

Facet 2: PSU item calibration IRT model fit and Data Collection	
1. Describe model fit approach followed in PSU field test 2. Describe characteristics of data collection design	
3. Follow up questions (if needed and if applicable)	Rating
a. What IRT model(s) fit indices are used? • Are there indicators of model fit used in PSU FT calibration? • Is there a record of misfitting items? • Is there available documentation describing process to reconcile model lack-of fit?	C (3.9)
b. What types of applicants' parent populations are targeted during the sampling design? • What are relevant variables involved in sampling? • What type of FT sampling schema is followed? • What process is followed to evaluate implementation of the FT sampling plan?	C (3.9)

Table 103 shows a summary evaluation for calibration with synthetic data set. The run of independent item calibrations was performed utilizing BILOG 3.11 software and synthetic data. Three thousand simulees with ability distributed normally with a mean of 0 and a unit standard deviation took a 50 multiple-choice item test with known item parameters. A vector of item responses for each examinee was computed utilizing information from each simulee's ability and item parameters. Probabilities of correct responses were then transformed to vectors of valid scores utilizing an IRT model allowing for a guessing parameter and standard psychometric procedures described elsewhere (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). Appendix B provides a summary of descriptive item statistics, model fit, and simulees' ability distribution on a simulated data set.

Recovery of item parameters showed a high degree of correlation in the two calibration groups. Correlations ranged from 0.985 to 0.996. Figure 37 shows a scatter plot for IRT difficulty parameter estimates. (Note: the process took a couple of dry-runs before DEMRE finalized their data cards for the run. Along the process, the evaluation team provided improvement feedback to finalize data cards.)

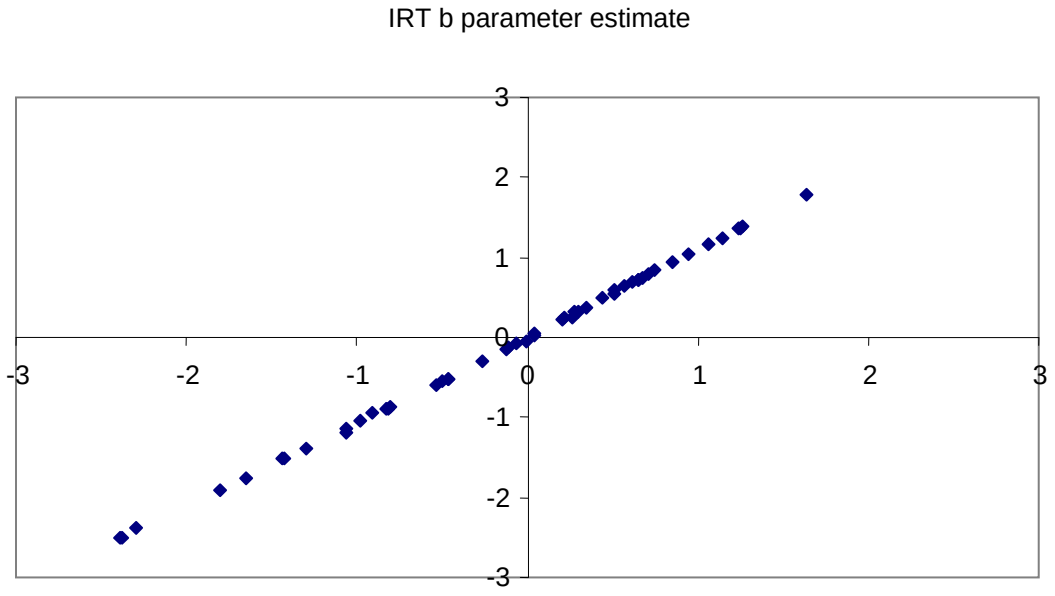


Figure 37: IRT Estimated Difficulty by Calibration Group (X=Evaluation Team; Y=DEMRE)

Table 103: Summary Evaluation of IRT Calibration and Equating (FACET 3)

Facet 3: IRT item calibration with Synthetic data set	
1. Given a simulated data set with known item parameters, use syntax and produce calibration results	
2. Follow up questions (if needed and if applicable)	Rating
a. Does independent calibrations with synthetic yield similar results?	E

Table 104 and Table 105 show summary evaluations of IRT PSU test equating processes and quality control processes. In Chile, the documentation of item response theory approaches to PSU test equating could be implemented in the near future. However,

currently PSU test scores are reported on a score scale with a mean of 500 points and a standard deviation of 110 points. Maintenance of the scale is an important endeavor in high-stakes testing, and this is typically performed through equating methodology and data collection designs. Furthermore, PSU test scores from a given year of administration are not equated to a common scale. As a result, a score of 650 points from the 2011 administration may not be comparable to a score of 650 points from the 2012 administration. The lack of year-to-year equating could hinder the validity of PSU score use for more than a single year, especially because PSU test scores can be used for up to two consecutive admission processes. There may be other relevant negative effects impacting the use of PSU test scores for (1) assigning social benefits (e.g., scholarships) and (2) performing research involving longitudinal changes of test scores, for example.

The greatest concern on the IRT equating for the PSU tests is the lack of a mechanism put forward for maintaining the PSU scale across years. As mentioned before, in Chile, PSU test performance is reported with the PSU scale but no equating takes place to maintain the scale across admission years. This is a serious issue that must be attended to ensure Chile university admissions tests are fair to test takers and to ensure valid comparisons of test scores across test administrations.

The anchor items are essentially useless since they do not fulfill their purpose. Even at the design stage, the anchor set shows multiple flaws. Anchor set content and statistical characteristics are deficient in light of the international standards. The anchors are too short and under representing the total test content. The implementation of such anchor specifications, if pursued, would render bias equating results and large equating error. Also the design of the anchor sets shows lack of awareness information about the actions taken, at least from a design perspective, to reduce context effects (e.g., retaining item position of anchor items) on performance of anchor items and processes to perform checks on anchor item parameter drift. To minimize context effects, anchor items should retain their position on the tests. The lack of plans for incorporating screening of item parameter drift presents potential threats to equating accuracy.

When moving toward incorporating PSU test equating to the processes, an anchor definition is of crucial importance, particularly when relying on data collection designs involving non-equivalent groups (Kolen & Brennan, 2004; Kolen 2006).

Table 104: Summary Evaluation of IRT Calibration and Equating (FACET 4)

Facet 4: PSU IRT Test Equating Methodology	
1. Describe IRT test equating methodology followed for PSU.	
2. Follow up questions (if needed and if applicable)	Rating
a. What IRT test equating methodology is followed for PSU • Type of data collection design • Type of IRT equating model • Process to implement PSU test equating • Evaluation criteria for PSU test equating • Types of scale transformation functions	A (3.9, 4.11, 4.13)
b. What are psychometric characteristics of set of anchor items for PSU test equating? • Anchor item set content and psychometric characteristics? • Anchor item sequencing in the test? • Anchor set role (e.g., internal and external)? • Anchor set stability check?	B (3.9, 4.13)

Table 105: Summary Evaluation of IRT Calibration and Equating (FACET 5)

Facet 5: PSU IRT Test Equating Quality Control	
1. Describe quality control procedures for PSU IRT test equating	
2. Follow up questions (if needed and if applicable)	Rating
a. What quality control procedures are followed for PSU IRT test equating (e.g., independent replication, equating specifications, dry-run)? b. What are qualifications of quality control staff for PSU IRT equating? c. What is the schedule for quality control activities to take place as part of PSU IRT equating?	D*

***A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

The international PSU evaluation team considers PSU equating to be below international standards. The documentation for PSU equating is not only misleading but also incomplete and inaccurate in some areas. For a national entrance exam with high stakes for thousands of applicants, the technical adequacy is insufficient, which means that erroneous outcomes (i.e., decisions) may occur. The evaluation team proposes a refocus of efforts that would address the following recommended improvements.

1. In order to replenish the item bank as new tests are created each year, newly developed items must be field tested and equated onto the scale of the original form. Once the field test items are administered, it is necessary to place their item parameters onto the same scale as the original form of the test in order to enable pre-equating during the test assembly process. Calibration of field test item parameters can be performed with approaches reviewed by Kolen & Brennan (2004).
2. In order to retain scale properties and allow comparability of test scores between years of test administrations, newly administered PSU tests need to be equated to compensate for differences in difficulty. A statistical equating simply establishes the relationship between two test forms. Typically, this is accomplished through the use of a common element across test administrations—either common persons or common items. In some cases, where appropriate, an assumption may be made that two separate groups taking two separate test forms are randomly equivalent. In most university admissions test contexts—where the goal is equating test forms from year to year—a common persons design is not typically feasible. Kolen & Brennan (2004) have an extensive treatment of alternative approaches to conduct test equating that can be consulted. It is important to emphasize that test equating is not the solution to test construction issues. The test construction process aims to develop a test form that is equivalent in content and difficulty to other forms administered in previous years. Equating is a tool that compensates for differences in test difficulties that could not have been controlled during test construction.
3. We recommend that the PSU equates test forms across test administrations. The lack of equated scores undercuts the ability to develop assessments that are fair to test takers. Fairness could be at stake when students taking PSU test on year 1 are advantaged with respect to those who took another PSU test on a subsequent year. For an assessment to be considered fair, test scores should not depend on the particular test form taken. In Chile, PSU test scores can be utilized up to two

consecutive years as part of the admission process. Equivalency of the PSU scores between forms is a necessary condition to support such an emergent use.

4. The design of the anchor set should comply to international standards. The design should describe targets of content coverage and psychometric representation of the anchor set in such a way that the anchor set can be seen as a mini-version of the total test. The design should describe measures to control for content effects and potential drift of item anchors.
5. The 2PL model is currently being used for item analysis in the PSU program without a rationale. If this model is to be used in the future for item analysis or for additional purposes, the evaluation team strongly recommends following international practice and validating the adequacy of its use over the typically used alternatives of the Rasch or 3PL models. Similarly, the evaluation team recommends developing documentation of item calibrations that are available to staff who participate in the calibration of items. Such documentations could be used to train staff on PSU item calibration processes that have been approved by DEMRE and the technical advisory committee (CTA).

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

DEMRE. (2011a). *Directrices psicométricas para el análisis de ítemes PSU.* Santiago: Universidad de Chile.

DEMRE. (2011b). *Procedimientos para determinar la muestra para el pretest*. Santiago: Universidad de Chile.

DEMRE. (2011c). *Criterios para la selección de preguntas de anclaje en ensamblaje de pruebas experimentales 2011* (admisión 2012). Santiago: Universidad de Chile.

Kolen, M. (2006). Scaling and norming. In R. L. Brennan (Ed.) *Educational measurement* (4th ed. pp. 155–186). Westpoint, CT: American Council on Education and Praeger Publications.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Objective 1.4. Evaluation of software and processes utilized for statistical analysis and item bank

DEMRE has a structure of data processing for statistical analyses and item banking of PSU test items that branches into two components. One branch, the DEMRE Unit of Informatics, has covered computation of CTT item analyses such as item difficulty and item discrimination utilizing standard database processes. The unit relies on item analysis software that computes CTT item statistics and populates such results in the item bank. The process is followed for both PSU pilot and operational administrations. The process takes the PSU scored data files (either pilot or operational) as inputs and reads them into the system for item analysis. Outputs of the data processing efforts are read into item bank automatically.

The other branch, the DEMRE Unit of Research, has covered analyses of differential item function (DIF) and item response theory (IRT) calibrations for both pilot and operational administrations of the PSU tests, starting from the 2006 admissions process. In previous years, statistical analyses were grounded solely on the basis of CTT, and the data processing was implemented through the unit of informatics item analysis module. DEMRE's Unit of Research performs DIF and IRT item calibration utilizing free and commercially available software, respectively. The data processing activity consists on receiving the scored data sets (either pilot or operational) and reading them into the corresponding piece of software. Results of the DIF and IRT item calibration efforts are outputted into excel files to be read into the item bank. The item bank software has the built-in capability to read such files as part of the loading of the IRT item calibration and DIF results.

The evaluation team met with relevant stakeholders from DEMRE on March 26, 2012, to fine-tune information concerning the processes and software used for statistical analyses and the maintenance of the item bank. The purpose of meeting was to gain a deeper understanding of the process and the flow of information among units processing data and loading data into the item bank. The meeting covered the following aspects:

- Gain details on DEMRE's processes
- Broaden understanding of the flow of data transfer and outputs from and to DEMRE units
- Fine-tune information on scope of statistics loaded into the item bank
- Fine-tune information on process to update statistics in the item bank
- Expand the understanding of the functionality of the item bank (capacity and flexibility to receive item statistics)

The meeting was held within DEMRE offices following an agreed-upon schedule for the visit. The following DEMRE staff participated in the interviews:

- Head of research unit and his team
- General coordinator
- Head of admissions process
- Director of DEMRE

Demographic survey and feedback information were collected from participants. The overall ratings of meeting preparation, quality of facilities, and readiness of interviewers indicated a high degree of satisfaction among interviewees with process that the evaluation team followed in the working session.

The following subsection contains the results of the evaluation for Objective 1.4.

GENERAL DESCRIPTION
Software includes:

Item Bank Database
Oracle 9i client/server database hosts the database that stores all items, statistics, and other associated data. The database is maintained by the DEMRE Information Technology Unit. Access to item data and item text is controlled according to the role of the user—only members of the Test Construction Unit have access to the text of the items and the item keys.

BILOG 3.11
BILOG 3.11 is used to generate parameters for the items using the 2-parameter logistic model. These parameters are used by the item bank database to display item characteristic and information curves.

DIFAS 4.0 (Penfield, 2007)
DIFAS reports the following statistical measures of differential item functioning (DIF) for dichotomous items:

- Mantel-Haenszel Chi-square
- Mantel-Haenszel common log-odds ratio
- Mantel-Haenszel estimated standard error
- Standardized Mantel-Haenszel common log-odds ratio
- Breslow-Day test of trend in odds ratio heterogeneity

In addition, two heuristic schemes for flagging items as possibly exhibiting DIF are available within the program:

- Combined Decision Rule (Penfield, 2003)
- ETS Categorization Scheme (Zieky, 1993)

SAS
Custom SAS code is used for the equating of the Science tests.

The international evaluation team relied on the following set of standards for quality control in scoring, test analyses and reporting of test scores from the International Test Commission (ITC, 2012) and the *Standards for Educational and Psychological Measurement* (AERA, APA, & NCME, 1999).

Standard 3.4.1.1.

> Use a reliable process to perform item analysis, and make sure that the programs have adequate technical documentation. (ITC, p. 14)

Standard 3.1

> Test and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development. (AERA, APA, & NCME, 1999, p. 43)

(Note: The proper interpretation of the above set of standards depends on professional judgments of the evaluators; the standards are intended to guide professional judgment, not to cancel the need for it.)

EVALUATION

The item bank database seems to be well designed with respect to the security of the items. The limitations placed on users minimize the possibility of security breaches for both operational items and forms. For example, only test authors can view items, authors can only view items associated with their subject areas, hardware-based keys are required for access to item images, psychometricians cannot view items or keys, and IT technicians, including database administrators, cannot view items or keys. There are options for authorized users to export the item images to *.DOC files that are saved on their local machines, so security of the tests should include ensuring the security of those authorized users' computers. These measures might include such things as encryption of those machines' hard disks, requiring that screen savers with passwords be enabled, and limiting networking of the machines to internal LANs only.

Version control of item images is present but seems weak based on the available documentation. Only authorized users can make changes, but it doesn't appear that these changes are tracked in any fashion or that previous versions of the items are retained. Items are locked at certain stages in the item development and usage cycle.

There is no documentation of version control of item statistics beyond an item status indicator that shows how many times an item has been used operationally. It appears that only the statistics from the most recent administration are retained in the database.

BILOG 3.11 and DIFAS pieces of software both use well-researched and recognized statistical procedures to estimate IRT item parameters and Mantel-Haenszel DIF statistics, respectively. BILOG 3.11 is limited to dichotomous items which is the item format currently in use for the PSU tests. As we have stated before, DEMRE relies on the 3.11 version of the BILOG 3.11 software. The evaluation team recommends developing an updating plan for software version updates. The BILOG 3.11 software may need to be replaced by other software, depending on future decisions. If in the future it is decided to include polytomous items on the PSU test, BILOG 3.11 would fall short when handling this item format type. Commercially available software that allows for polytomous items (e.g., MULTILOG) can be considered and evaluated. Likewise, if in the future it is decided to include IRT equating while allowing for the estimation of item parameters for more than one group, the BILOG 3.11 software should be replaced by software that allows for such kind of analyses (e.g., BILOG MG).

DIFAS accommodates multipoint items but only provides statistical measures of DIF without heuristic flags for such items. The item bank would need to be extended to provide such flags should multipoint items be added to the tests at some point in the future. In addition, the evaluation team recommends developing a plan for the software version update. (Note: More information on challenges when involving multiple DIF methodologies can be found as part of Objective 1.1.g. of this evaluation).

SAS is a robust system and is well suited for use for the Science equating analyses. The SAS code itself has sufficient documentation, and appropriate SAS PROCs are being used.

Table 106 shows a summary of software tools in the PSU university admissions program. Under column labeled "Rating," the table shows the judgment of the analytical evaluation and lists within parentheses the professional standards not met.

Coding Schema

A= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least three of the international assessment standards identified.

B= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least two of the international assessment standards identified.

C= Information present which does not comply with the minimum expectations (neither clear nor pertinent), and strays away from at least one of the international assessment standards identified.

D= Information not described in the documents nor was found when asked about during the interview.

E= Information present complying with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards.

F= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes an additional improvement element that has been carried out in the development of the PSU.

G= Information present and in compliance with the minimum expectations (clear and pertinent). This element complies with the identified international assessment standards. As well, includes two or more additional improvement elements that have been carried out in the development of the PSU.

Table 106: PSU Software Tools

Facet 1. Statistics/variables contained in the item bank	Rating
a. User accessibility	E
b. Data location	E
c. Data security	E
d. Backup procedures	D*
e. Transparency	D*
Facet 2. Maintenance processes (updates of existing variables and addition of new variables)	**Rating**
a. Version control (item images)	C (3.4.1.1)
b. Version control (item statistics)	D*
c. User Access	F
d. Revisions by database manager	E
e. Process documentation	E
f. Monitoring of metadata	D*
g. Research-based documentation	E
Facet 3. Software functionality (current capacity and extensibility)	**Rating**
a. Criteria used to select software	D*
b. Criteria used to select software version	D*
c. Flexibility of software	C (3.4.1.1)
d. Version control (software)	E
e. Data storage capacity	E
f. Open source software	D*

*A rating of D= Information was not described in the documents nor was it found when asked about during the interview.

RECOMMENDATIONS

The international PSU evaluation team considers the set of software tools available for analysis of the PSU university admissions testing program to be below international standards. The evaluation indicates that though there may be just enough automation *within* a functional group, there is not enough *among* functional groups. For a national entrance exam with just a single operational administration of six assessments, the processing environment may be tolerable. However, the PSU testing program would be challenged were it called upon to allow for multiple test administrations, something which occurs regularly in many university admissions programs internationally. The evaluation team proposes a refocus of efforts that would address the following recommended improvements.

1. The current systems seem to be somewhat disjointed, with much manual manipulation of item and test data required. One of the verification steps during test construction is to check the item codes to verify that they exist in the database, implying that the test author must type in the item codes manually. This is a place where an error can occur if the test author mistypes an item code and the mistyped code happens to match another (unwanted) item in the bank. The SAS code used to implement the Science test equating appears to require manual editing for each successive year. There are many references to "importing" and "exporting" data to and from the database; however, to the extent that these functions require manual manipulation, these are steps in the process where errors can occur. It is recommended that common processes be automated as much as possible and that analyses be standardized to eliminate or reduce the amount of manual intervention required.
2. Versioning of both item images and item statistics can be improved. The ID of users making modifications to items should be tracked. In addition, previous versions should be retained, both to provide a historical record of changes made to an item and also as a safeguard to allow reversion to an earlier version if needed.
3. Statistics from all administrations of an item should be retained, and users should be able to view them together in chronological order. Large changes in item statistics from one administration to another can indicate such things as item exposure, printing errors, or other problems. The documentation indicates that key-check analyses are performed for field test (tryout) items; key-check analyses are recommended for all items to control for unrecorded changes in items, detection of printing or production errors, errors in the importing and exporting of data to and from the database or other unforeseen errors.
4. BILOG 3.11 software can be run in batch mode, and command files can be produced automatically by custom-written software or via SAS programs, both of which are recommended. BILOG is a rather old program, and newer software might bring benefits. If using IRT for equating is possible in the future, it might be worthwhile to upgrade to newer software.
5. DIFAS does not allow for the use of command files and therefore cannot be run in batch mode. Mantel-Haenszel analyses can be easily coded in custom software or can be run in SAS. Replacing DIFAS with a solution that can be more easily automated would reduce the need for the labor involved in running DIFAS, saving the results and then importing them into the database.
6. The general process for producing the normalized scores was described in the documents available for this review, but the specific procedures and software used to accomplish these analyses were not. In general, the preceding recommendations

also apply to the processes and software used for the derivation of the normalized scores. To the maximum extent possible, these processes should be automated so that they can run without manual user intervention, and software amenable to running in batch mode (such as SAS) should be used for these derivations.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

International Test Commission (2012). *ITC guidelines for quality control in scoring, test analysis, and reporting test scores*. ITC: Author.

Penfield, R. D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of non-uniform DIF. *Alberta Journal of Educational Research, 49*, 231-243.

Penfield, R. D. (2007). *DIFAS 4.0 user's manual*. Retrieved from http://www.education.miami.edu/facultysites/penfield/index.html

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

Objective 1.5. Evaluation of the delivery process and of the clarity of information regarding those examined and the different users of the admissions system

Internationally, score reporting is a critical component of any testing endeavor, which in many instances has been neglected. From stakeholders' perspective, score reporting sometimes has been misunderstood, misinterpreted, and potentially confused stakeholders.

The *Standards for Educational and Psychological Testing* (APA, AERA, NCME, 1999) brings a number of references to score reporting and interpretative materials.

Standard 5.10

> When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, and the precision of the scores, common misinterpretations of test scores, and how scores will be used. (p. 65)

Standard 6.3

> The rationale for the test, recommended uses of the test, support or such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified. (p. 68)

Standard 8.8

> When score reporting includes assigning individuals to categories, the categories should be chosen carefully and described precisely. The least stigmatizing labels, consistent with accurate representation, should always be assigned. (p. 88)

Good score reporting communicates the results in a manner that is tailored to the characteristics of audience (Goodman & Hambleton, 2004).The purpose of this study is to examine PSU score reports and processes as well as to provide stakeholders' opinions on needs for improvement.

METHODOLOGY

Following international practices, the evaluation team interviewed groups of intended audiences of the reports to gather information on whether these stakeholders understood the information contained in the current reports and if they thought there was any need for improvement of the reports and processes. The interview questions were targeted to three distinct groups based on the information they currently receive from DEMRE: students, high school teachers and university admissions officers.

Gaining access to these stakeholders was a challenging activity. Because the academic calendar involved final tests and winter vacation, interviews with groups of interest took a considerable amount of time to complete. Table 107 shows expected and actual number of interview by group of stakeholders.

Table 107: Number of Interviews with Specified Stakeholder Groups

Stakeholder groups (Score reporting)	Expected number of interviews	Performed number of interviews
Students	10	8
High school teachers	10	6
Admission officers	12	7

The evaluation team could not fulfill the targeted number of interviews because of a particularly low response rate from the stakeholders contacted after several attempts and by various means (e.g., letters, email, phone calls).

There is no single universal rule available for practitioners when selecting the number of stakeholders to involve in score reporting analyses. Practitioners often seek to arrive at groups of participants that capture diversity of score report stakeholders and to hear from them their experiences and recommendations. A tradeoff between quantity and quality was encountered along the evaluation of PSU score reports. The international evaluation team review recent relevant literature on score reporting to document practical sample sizes used in this type of studies and to help with the conundrum.

Table 108 summarizes the findings. The purpose of the table is to depict reasonable number of stakeholders considered for score reporting activities. The table shows numbers of stakeholders ranging from two to 16 stakeholders. The lower bound of n-counts is from a study where two measurement professionals inspected student test score reports. The upper bound of n-counts is from a study where 16 professional (high school teachers and college faculty) participated in interviews geared toward score reporting on teacher certifications. The table also shows a tendency to involve fewer stakeholders as long as they possess academic credentials, but the number may increase by a notch for achieving geographical coverage. For example, in Klesch study six assessment specialists participated in score reporting meeting/interviews whereas in Zenisky, Hambleton, and Sireci (2009) 15 individuals from state and district education offices and policy makers participated in the study.

Table 108: Summary of Stakeholders in Recent Score Reporting Studies

Authors	Context	Stakeholders
Zenisky, Hambleton, and Sireci (2009)	NAEP score reporting practices	Fifteen individuals representing state and district education offices and policy makers
Goodman and Hambleton (2004)	Students test score reports	Two people
Klesch (2010)	Teacher certification score reports	Sixteen educators from public schools and colleges through the U.S. Six assessment specialists Ten university doctoral students

Trends captured from professional literature on score reporting, the evaluation team delved more from mixed-methods literature. The intention for this additional effort was to learn whether the typical sample sizes often involved in rigorous score reporting studies correlates with number of stakeholders involved in rigorous mix-methodology studies. For example, in Onwuegbuzie and Combs' (2009) paper, seventeen doctoral students were interviewed and asked about coping mechanisms toward a content related anxiety. In their sequential methodology the authors involved a survey questionnaire and 115 graduate students. Because of its in-depth nature and face-to-face interaction, the detailed responses provided by a limited number of stakeholders served the basic purpose of this investigation.

The international evaluation team feels confident on the number of stakeholders participating in score reporting interviews. The numbers not only mimics number of stakeholders from rigorous studies but also show correlation with numbers often encountered in field studies with mixed methods. The numbers of stakeholders were chosen from a list of potential stakeholders from Chile's Ministry of Education.

The evaluation team would like to advise readers of these interviews that they provide a partial picture of the full spectrum of possible responses that could be drawn from stakeholders concerning the PSU score reports.

INSTRUMENTS

The follow pages show the questions presented to each group of stakeholders during their interviews.

Pauta Entrevista: ***Estudiantes***
Reporte de entrega de resultados PSU

Instrucciones: Familiarícese con el *reporte de entrega de resultados PSU*. Prepárese para discutir las siguientes preguntas:

1. ¿Para cual proceso de admisión fue elaborado el *reporte de entrega de resultados PSU*?

2. ¿Cuántos postulantes se encuentran resumidos en este reporte de entrega de resultados PSU?

3. Respecto de la prueba de Lenguaje, ¿en qué medida el reporte le ayuda a interpretar el puntaje PSU de 727?

4. Respecto de la prueba de Lenguaje, ¿en qué medida la información en el reporte le ayuda a identificar el número de preguntas correctas que corresponden al puntaje PSU de 727?

5. A partir de la información presentada, ¿en qué medida la información en el reporte le ayuda a entender la precisión alrededor del puntaje PSU en Lenguaje de 727?

6. De compararse el puntaje PSU en Lenguaje de 727 con otro puntaje PSU en Lenguaje de 726, ¿en qué medida la información en el reporte le ayuda a interpretar la diferencia entre los dos puntajes?

7. Escriba un listado de todos los usos que le da al *Reporte de entrega de resultados PSU*. Al finalizar el listado, elabore un ranking de importancia de usos, identificado con el numero 1 al uso de mayor importancia, el 2 para el siguiente uso en importancia, y así sucesivamente.

8. Describa en qué grado el reporte de entrega de resultados PSU satisface sus necesidades de información. (Niveles de grados: 100%, 90%, 80%, etc.). Por favor fundamente su respuesta.

9. Comparta alguna opinión y/o sugerencia que usted desee expresar con respecto al proceso que se sigue para hacer llegar el *reporte de entrega de resultados PSU*.

Pauta de Entrevista: ***Estudiantes***
Reporte de admisión a las universidades chilenas

Instrucciones: Familiarícese con el *reporte de admisión a las universidades chilenas*. Prepárese para discutir las siguientes preguntas:

1. ¿En qué medida el reporte de admisión a las universidades chilenas describen el proceso para el cálculo del puntaje ponderado para la admisión?

2. Para un puntaje ponderado de 724.30, ¿de qué manera el reporte le ayuda a interpretar dicho puntaje?

3. A partir de la información presentada, ¿de qué manera el reporte de admisión le ayuda a conocer las ponderaciones que son utilizadas en el cálculo del puntaje ponderado por las universidades (carreras) a las que postulo?

4. A partir de la información en el reporte de admisión, me puede decir, ¿Cuál es el cupo disponible en la universidad que el postulante eligió como su segunda opción?

5. Escriba un listado de todos los usos que le da al *Reporte de admisión a las universidades chilenas*. Al finalizar el listado, elabore un ranking de importancia de dichos usos, identificado con el numero 1 al uso de mayor importancia, el 2 para el siguiente uso en importancia, y así sucesivamente.

6. Describa en qué grado el *Reporte de admisión a las universidades chilenas* satisface sus necesidades de información. (Niveles de grados: 100%, 90%, 80%, etc.). Por favor fundamente su respuesta.

7. Comparta alguna opinión y/o sugerencia que usted desee expresar con respecto al proceso que se sigue para hacer llegar el *Reporte de admisión a las universidades chilenas*.

Pauta Entrevista: **Profesores de enseñanza media**
Informe estadístico de resultados de la PSU

Instrucciones: Basándose en su familiaridad con el *Informe estadístico de resultados de la PSU* disponible para su colegio, por favor prepárese para discutir las siguientes preguntas:

1. ¿Cómo interpreta el puntaje PSU?

2. ¿Cómo interpreta el puntaje NEM?

3. Escriba un listado de todos los usos que le da al informe estadístico. Al finalizar el listado, elabore un ranking de importancia de usos, identificado con el numero 1 al uso de mayor importancia, el 2 para el siguiente uso en importancia, y así sucesivamente.

4. A partir de la información en el reporte, ¿se podría distinguir entre los estudiantes seleccionados y los estudiantes no seleccionados? ¿Qué parte del informe estadístico le permite llevar a cabo dicha identificación?

5. Describa en qué grado el capítulo 1 (resultados en puntajes estandarizados) satisface sus necesidades de información. (Niveles de grados: 100%, 90%, 80%, etc.). Por favor fundamente su respuesta.

6. Describa en qué grado el capítulo II (análisis de resultados por área temática y habilidad) satisface sus necesidades de información. (Niveles de grados: 100%, 90%, 80%, etc.). Por favor fundamente su respuesta.

7. Describa en qué grado el capítulo III (análisis comparativo por dependencia, nivel de comuna, región y país) satisface sus necesidades de información. (Niveles de grados: 100%, 90%, 80%, etc.). Por favor fundamente su respuesta.

8. Comparta alguna opinión y/o sugerencia que usted desee expresar con respecto al proceso que se sigue para hacer llegar el *Informe estadístico de resultados de la PSU*.

Pauta Entrevista: ***Encargados de Admisión de las Universidades***
Procesos e información diseminada por el DEMRE a las universidades

Instrucciones: Prepárese para discutir las siguientes preguntas en las siguientes dos secciones: Procesos de admisión basados en la PSU y procesos paralelos de admisión.

I. PROCESO DE ADMISSION BASADO EN LA PSU

9. Describa el proceso que sigue el DEMRE y su universidad para comunicar resultados del de selección. En su respuesta mencione las etapas del proceso, la descripción de cada etapa, y el objetivo que se persigue. Finalmente proporcione su grado de satisfacción con cada etapa del proceso utilizando la siguiente escala:

    a. 5 = Totalmente Satisfecho
    b. 4 = Muy Satisfecho
    c. 3 = Satisfecho
    d. 2 = Muy Insatisfecho
    e. 1 = Totalmente Insatisfecho
    f. N/O = Desconozco el proceso

Nombre de la etapa	Descripción de la etapa	Objetivo de la etapa	Grado de satisfacción

10. Comparta alguna opinión y/o sugerencia que usted desea expresar con respecto al proceso que sigue el DEMRE para comunicar resultados del proceso de selección a su universidad.

_____
_____
_____
_____
_____
_____
_____
_____

11. Describa la información (informes, reportes, bases de datos, etc.) de los resultados de los procesos de selección que el DEMRE le proporciona a su universidad. En su respuesta proporcione el nombre de la información (p.ej., reporte tal cual, base de datos tal cual), la descripción de la información y el uso que le da a la información. Finalmente proporcione el ranking de importancia de la información recibida proporcionado el numero 1 a la información de mayor importancia, el 2 a la siguiente en importancia y así sucesivamente. Utilice N/O en el ranking cuando usted desconozca la información.

Nombre de la información	Descripción de la información	Usos que le da a la información	Ranking

12. Comparta alguna opinión y/o sugerencia que usted desea expresar con respecto a la información (por ejemplo, informes, reportes, bases de datos) que el DEMRE proporciona (o no proporciona) a su universidad.

_____
_____
_____
_____
_____
_____
_____
_____

II. PROCESOS PARALELOS DE ADMISION

- ¿Existen procesos paralelos de admisión en su universidad? ¿Cuál es el criterio de ingreso en dichos procesos paralelos? ¿Qué tipo de postulante tienen acceso a los procesos paralelos? ¿Cuál es la cuota máxima de postulantes para los procesos paralelos?

Nombre del proceso paralelo de admisión	Descripción del proceso paralelo de admisión	Criterio de ingreso	Características de los postulantes y cuotas

2. ¿En el caso de que su institución aplique pruebas especiales, cuál es el valor en relativo en calidad sobre lo que aporta la PSU y NEM? ¿Qué uso le dan a los puntajes en las pruebas especiales?

_____
_____
_____
_____
_____
_____
_____
_____

MUCHAS GRACIAS

RESULTS

There are two tables presenting summaries of the opinions from student stakeholder group. Each table contains the questions asked the students and summaries of the major points gathered from transcripts of the interviews.

Table 109 is a summary of the Students' responses to the *PSU Delivery Report Results*. The major overall theme that can be gleaned from interviews with the students is that while they understood the intent and basic purpose of the report, they were not able to gather much useful information from the reports. The students did not believe the report contained information that would help them answer any quantitative questions about scores in the report. Many of the students wanted more specific, diagnostic feedback from the report.

Table 109: Summary of Students' Responses to PSU Delivery Report Results

PSU Delivery Report Results
1. For what admissions process was the *PSU Delivery Report Results* developed? • All said university admissions process.
2. How many applicants are summarized in this report for deliverables PSU? • All said it was an individual report.
3. Regarding the language test, to what extent the report helps you interpret the PSU score of 727? • Most of the subjects did not understand the report. • A few subjects thought it might be a good score or good enough to be admitted.
4. Regarding the language test, to what extent the information in the report helps you identify the correct number of questions corresponding to PSU score of 727? • The subjects indicated the information cannot be found or that they could not determine the answer.
5. From the information provided, to what extent the information in the report helps you understand the precision around the PSU in Language score of 727? • The subjects indicated the information cannot be found or that they could not determine the answer.
6. To compare the score of 727 in Language PSU with another PSU in Language score of 726, to what extent the information in the report will help to interpret the difference between the two scores? • The subjects indicated the information cannot be found or that they could not determine the answer.
7. Write a list of all applications that gives the delivery of results Report PSU. At the end of the list, create a ranking of importance of uses identified by the number 1 to use the most important, 2 for the next use in importance, and so on. • For most subjects, the first use of the report is for university admissions information • For other subjects, the report is used for self-evaluation of skills do they know.

PSU Delivery Report Results
8. Describe to what extent the results report delivery PSU meets their information needs. (Degrees Levels: 100%, 90%, 80%, etc.). Please substantiate your answer. • A majority of subjects felt that the report did not meet all their needs but only gave the number of correct answers. • All subjects wanted more feedback on what they got right and what they got wrong. More information would be better.
9. Share any thoughts and/or suggestions you wish to express regarding the process followed to get the delivery report results PSU. • The subjects felt the process was alright because there were many different ways to obtain the results. • Some subjects want more feedback such as the questions they got right and the ones they got wrong.

Table 110 summarizes the reactions of the students to the *Chilean University Admissions Report*. Overall, the students did not feel the report gave enough if any information for them to be able to answer any quantitative questions from the report. Basically, the subjects felt the report was of little help and did not provide enough information.

Table 110: Summary of Students' Responses to Chilean University Admissions Report

Chilean University Admissions Report
1. To what extent the report of Chilean university admissions describe the process for calculating the weighted score for admission? • Subjects indicated report does not describe and give no information on the weighted score. • Only shows the scores.
2. For a weighted score of 724.30, in what way the report helps you interpret that score? • Nothing is in the report at all and it only gives the score and if on the waiting list.
3. From the information provided, in what way the intake report helps you know the weights that are used in the calculation of the weighted score universities (races) that apply? • Nothing is in the report at all and it only gives the score and if on the waiting list.
4. From the information in the report for admission, I may say, What is the space available in the university that the applicant chose as his second choice? • One subject said 108, two said 58, one said there were no openings, and the other did not know. None seemed to find the information to be able to answer the question easily.
5. Write a list of all applications that gives the Report Chilean university admissions. At the end of the list, create a ranking of importance of such uses, identified by the number 1 to use the most important, 2 for the next use in importance, and so on. • It really only relates whether a student is accepted, not accepted, or on the waiting list and not much else.

6. Describe to what extent the Report Chilean university admissions meet their information needs. (Degrees Levels: 100%, 90%, 80%, etc.). Please substantiate your answer. • Slightly different answers from the subjects but none felt they got the information they needed.
7. Share any thoughts and / or suggestions you wish to express regarding the process followed to get the Report Chilean university admissions. • All subjects expressed the report was of little help and did not provide enough information.

Table 111 is a summary of the Teachers' responses to the *PSU Statistical Reports*. Questions asking about uses and applications of the report yielded a variety of answers. These teachers felt the report could be used to project how the students might do in the future, while others believed it could be used to measure how well the students knew the content at the time of testing. Subsequent Questions asking the teachers to use the information from the *Reports* indicated a general difficulty finding the targeted information. Distributing the reports faster and more efficiently were two requests from the panel.

Table 111: Summary of High School Teachers' Responses to PSU Statistical Reports

PSU Statistical Reports
1. How do you interpret the PSU score? • Most subjects felt it assesses content and can be used for ranking student achievement. • Some subjects believe it can be used to project how students will do in the future.
2. How do you interpret the NEM score? • Varied answers from the subjects. Some believe it measures content, is an average of grades, or can be used to project future academic performance. • It can be used as a complement to ranking of students.
3. Write a list of all applications that gives the statistical report. At the end of the list, create a ranking of importance of uses identified by the number 1 to use the most important, 2 for the next use in importance, and so on. • A variety of applications were given by the subjects. Some used it for projecting how students will do in the future, some for comparing to other universities, and some in a diagnostic way to see what were the teachers or subjects that needed help. • Generalizing to how the student behaves.
4. From the information in the report, could we distinguish between students selected and unselected students? What part of the statistical report allows you to carry out such identification? • All subjects said no.
5. Describe what grade chapter 1 (results in standardized scores) satisfy their information needs. (Degrees Levels: 100%, 90%, 80%, etc.). Please substantiate your answer. • Only half the subjects answered this question. The half that did said 100%.
6. Describe to what extent Chapter II (analysis of results by subject area and skill) meet their information needs. (Degrees Levels: 100%, 90%, 80%, etc.). Please substantiate your answer.

• Only two subjects answered this question and had different answers. One indicated 100% and the other only 10% because students had different courses in high school.
7. Describe to what extent Chapter III (comparative analysis by dependency level of community, region and country) meet their information needs. (Degrees Levels: 100%, 90%, 80%, etc.). Please substantiate your answer. • Only two subjects answered this question. One felt it was at a country level and the other did not feel it was applicable because that high school did not have high PSU scores.
8. Share any thoughts and/or suggestions you wish to express regarding the process followed to get the statistical report of results of the PSU. • A more efficient manner to distribute the reports. Basically, DEMRE should deliver them. • Decrease the length of time used to deliver the reports.

The tables below present summaries of the opinions from university admissions officers. Each table contains the questions asked the Officers and summaries of the major and minor points gathered from transcripts of the interviews.

Table 112 is a summary of responses to the PSU Admissions procedure. This interview protocol included Likert-type responses to some questions and then ranking of uses and pieces of information. Overall, the officers indicated they were "satisfied" with the process that was used and for each specific stage of the process. While subjects gave different rankings to the importance of pieces of information from the reports, they expressed a desire to get the information sooner because of tight deadlines. One also mentioned that the information they were given is much better than before.

Table 112: Summary of Admissions Officers' Responses to PSU Admissions Procedure

PSU Admissions Procedure
1. Describe the process by which the university DEMRE and to communicate results of selection. In reply mentioned stages of the process, the description of each stage, and the aim pursued. Finally provide your level of satisfaction with each stage of the process using the following scale: **a. 5 = Completely satisfied** **b. 4 = Very Satisfied** **c. 3 = Satisfied** **d. 2 = Very Dissatisfied** **e. 1 = Completely Dissatisfied** **f. N/O = not know the process** • All subjects on typically gave a "3" to each stage of the process indicating they were satisfied.
2. Share any thoughts and/or suggestions you would like to express regarding the process followed by the DEMRE to communicate results of the selection process at their university. • Subjects thought more time should be scheduled for the selection and subsequent registration of students. • Subjects thought that the system used for getting information from DEMRE is cumbersome.

3. Describe the information (reports, reports, databases, etc.) of the results of the selection process that DEMRE gives your university. In response provide the name of the information (e.g., report as is, database as is), the description of information and use that information gives. Finally provide the ranking of importance of information received provided the number 1 to the most important information, the 2 to the next most important and so on. Use N/O in the ranking when you unfamiliar information. • First, the information on candidates and scores is received from DEMRE, after which the information from the institutions is uploaded, including which students are accepted, and then, finally, the vacancies are looked at. Information for special admissions is also compiled. Information on vacancies was rated important by at least one subject. • Each subject gave different rankings for each one of the pieces of information they used and received.
4. Share any thoughts and/or suggestions you wish to express regarding the information (e.g., reports, and databases) that DEMRE provided (or not provided) to your college. • Subjects would like the information to arrive sooner because of tight deadlines. More than one thought the information was much better than before.

Table 113 presents the responses of panelists who were asked two questions about the parallel admissions processes. The major non-PSU-related criteria included participation in athletics and a relationship with a university official.

Table 113: Summary of Admissions Officers' Responses to Parallel Admissions Processes

Parallel Admissions Processes
1. Are there parallel processes in university admissions? What is the criterion for admission for these parallel processes? What types of entrants have access to parallel processes? What is the maximum quota of candidates for parallel processes? • The main non-PSU criterion seemed to be having a university official as a relative. • Being an athlete or having a special admissions applicant was another way to get into university.
2. In case your institution administers special tests, what is the relative value of the weights applied to PSU and NEM scores? What use is given to the scores from special tests? • They are only used for outstanding athletes and for those that need to be brought up to a certain academic level before entering.

RECOMMENDATIONS

Our recommendations are grouped by stakeholder.

Students

PSU Delivery Report Results

1. Because students did not understand the PSU scale scores that were presented to them, we recommend that DEMRE provide additional interpretive information explaining the *PSU Delivery Report Results*.
2. We recommend that more information be provided regarding the areas of test takers' strengths and weakness on each PSU test, using the *PSU Statistical Reports* for educators as a model for a modified *PSU Delivery Report Results*. It is important because students can take PSU tests more than once and should have an opportunity to obtain feedback and improve their scores.

Postulation Scores

3. We recommend that general information should be provided explaining the weighting process and the cut scores used to create the postulation scores. In particular, DEMRE may want to consider providing information to students about particular universities and departments concerning the weights and cut scores.
4. We recommend that the reports be redesigned to make it easier for students to find information, such as the number of spaces available in university departments.

Educators

PSU Statistical Reports

5. We recommend that the information provided regarding the areas of test takers' strengths and weakness on each PSU test in the PSU Statistical Reports for educators be suspended until the results are carefully scrutinized to ensure the reliability and validity of such information.
6. Because educators indicated that they use the results of the PSU tests for purposes other than university admissions, we recommend that the *PSU Statistical Reports* explain what the intended uses are for the PSU tests and warn against the unintended uses.
7. Although the report contains a great deal of information, much of it quite valuable, it is difficult to find specific data to answer specific questions. For example, the report provides in the appendix the number of students admitted into university from their particular high school. However, educators were unable to locate this information. We recommend including a detailed table of contents that would improve the value of this report, making such information more readily available to educators.

University Administrators

PSU Admissions Procedure

8. We recommend reviewing the schedule for registration and admissions to facilitate the work of university administrators during the admissions process.

Parallel Admissions Processes

9. No recommendations are offered by the evaluation team.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220.

Klesh, H. (2010). *Score reporting in teacher certification testing: A review, design, and interview/focus group*. Doctoral dissertation. University of Massachusetts- Amherst.

Onwuegbuzie, A. & Combs, J. (2009). *The relationship between statistics anxiety and coping strategies among graduate students. A mixed research study*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Zenisky, A., Hambleton, R. & Sireci, S. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education*, *22*, 359-375.

Objective 2.1. Internal structure of PSU exams: goodness of fit of PSU test scores analyzed with item factor analysis and item response theory models

ABSTRACT

In developing tests and intended test uses, it is often useful to develop the conceptual definitions first and then to find ways to define them operationally. Cronbach and Meehl (1955) introduced the concept of construct validity to study whether test scores are sufficient to recover components highlighted in the conceptual definitions of the tests. The purpose of Objective 2.1 is to examine the internal item structure of the PSU test battery with item factor analytic and IRT frameworks utilizing PSU data from the 2012 admissions process. Item factor analysis was used to study the overall dimensionality of the PSU tests; the univariate IRT analysis (differential test functioning, or DTF) was used to study the invariance of tests structures across subpopulations.

Overall, the item factor analyses show that the PSU tests are essentially unidimensional. Most of the PSU tests show some evidence of differential test functioning, suggesting that the invariance of factor structure across subpopulations has only been partially achieved. The strongest evidence of DTF is seen for lower performing Technical-Professional students relative to higher performing Scientific-Humanistic students on the Science-Biology, Biology (common), Science-Chemistry, Language and Communication, and Mathematics tests. Regarding curricular branch, differences between lower performing Technical-Professional students relative to higher performing Scientific-Humanistic students was observed for the Science–Biology, Science-Biology (Common), Science-Chemistry, Language and Communication and Mathematics tests. Differences favor the latter group (Scientific-Humanistic).

For type of financing, differences between Private and Municipal (favoring Private schools) were found for the Mathematics, Language and Communication, History and Social Sciences and Science-Biology tests. For the comparison between Subsidized and Municipal schools, DTF differences were found for Mathematics. The difference favored the Subsidized group.

Finally, for the SES, the Mathematics test and the Language and Communication test showed DFT for most comparisons. Differences consistently favored the higher SES in the pairwise comparison.

INTRODUCTION

There are several approaches to documenting construct validity for an assessment in the psychometric literature, ranging from analyses of internal structure to analyses of validity generalization (AERA, APA, & NCME, 1999; Kane, 2006). The study of the internal structure of university admissions testing has low prevalence (Arce-Ferrer & Corral-Verdugo, 2002). Validity research on the overall dimensionality of the PSU tests and its invariance across subpopulations is scarce. There has been no study conducted to date in which item factor analysis and item response theory models were used to study the properties of the PSU test scores. In addition, there has been no study of invariance of a unidimensional structure between PSU forms. DEMRE uses two forms in their operational admission of the PSU, where Forms 1 and 2 contain the same items but are ordered in different ways.

The purpose of Objective 2.1 is to examine the internal item structure of the PSU test battery with item factor analytic and item response theory (IRT) frameworks. As agreed-upon during the goal clarification meeting, the analyses are performed with operational data

from the 2012 admissions process. This research explores the fit of models for major subpopulations. The background variables used throughout this report are listed below:

- Gender: Male or Female
- Regions: North (codes 1, 2, 3, 4, 15), Central (5, 13 [Metro]) o South (6, 7, 8, 9, 10, 11, 12, 14)
- Socio-economic status: Five quintiles of the SES variable—Quintile A defines the Lower group; Quintile B defines the Below Average group; Quintile C defines the Average group; Quintile D defines the Above Average group; and Quintile E defines the Upper group. SES was computed utilizing information from applicants' family income and parental education.
- Curricular Branch: Scientific-Humanistic or Technical-Professional
- Type of Financing: Private, Subsidize or Municipal

METHODOLOGY

Factor analysis is a statistical framework for investigating how and to what extent sets of observable variables relate to their underlying set of latent variables (McDonald, 1985). By identifying groups of correlated variables, factor analysis enables researchers to determine the number of latent variables (i.e., factors) underlying a set of observable variables. The factor analytic framework is flexible and accommodates exploration and confirmation of relationships between observed measures and their latent dimensions. In the exploratory mode, factor analytic endeavors look to understand links between observed variables and latent variables, since such relationships are uncertain and in many instances unknown. In contrast, the confirmatory mode of factor analyses seeks to corroborate initial understanding of how and to what extent observed variables are linked to their underlying latent dimensions.

Item factor analysis is one of the many applications of the factor analytic framework and seeks to understand how and to what extent observed item responses are linked to their underlying latent variables (Lord & Novick, 1968). Within the factor analysis framework, several approaches are available for item factor analysis of binary data (Bock, Gibbons, & Muraki, 1988; Joreskog, 1994; Muthén, 1978). The approaches rely on a factor-analyzing, item-by-item correlation matrix in which the dichotomous item responses are assumed to arise from their discrete set of latent continuous values.

To the degree that associations among items can be accounted for a single latent variable, item factor analyses and item response theory are essentially common. To paraphrase Muthén & Muthén, citing Baker & Kim (2004) and du Toit (2003), *confirmatory factor analysis* (CFA) is used to study relationships between a set of observed variables or item responses and a set of continuous latent variables or factors. When the item responses are categorical, CFA is also referred to as item response theory (IRT) analysis (Muthén & Muthén, 2007, Chapter 5, p. 49). We exploited the commonality between the two frameworks to study factor invariance across subpopulations with univariate IRT application known as differential test functioning (DTF).

Investigations of DTF involve gathering empirical evidence to show how members of a minority or focal group perform relative to a majority or reference group. This marshalling of empirical evidence is necessary, but it is alone insufficient to conclude that bias is present. A conclusion that bias is present necessarily involves making an inference that extends beyond any empirical analysis.

The logic behind using DTF to study bias at the test level is analogous to the logic of studying bias at the item level. To distinguish the empirical evidence from the inference, the term differential item functioning (DIF) rather than bias is used to describe the empirical evidence obtained in such an investigation. From the psychometric perspective, "an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, et. al., 1991, p. 109-110). In the absence of DIF, and except for the presence of random error, the probability of a response to a particular item on a test *Y* is determined solely by ability θ. If *Y* is determined solely by θ, then for any two groups *R* and *F*:

$$f(Y \mid \theta, G = F) = f(Y \mid \theta, G = R), \tag{1}$$

where the probability distribution of *Y* conditional on θ is denoted *f(Y | θ)* and grouping variable *G* is used to identify focal group *F* and reference group *R*. This equation shows that, in the absence of DIF, the conditional probability distribution for *Y* is independent of group membership *G*. Without losing generality, the above definition also applies to test scores. In the absence of DTF, the true-score for ability θ will be the same regardless of group membership. However, if DTF exists, then the expected proportion correct at θ will be different in the focal and reference groups. The larger the difference $D = T_F - T_R$, the greater the differential functioning of the test.

A measure of DTF at the person level may be defined as $D^2 = (T_F - T_R)^2$. Integrating over the focal group distribution for θ yields an unsigned DTF index:

$$\mathrm{DTF} = \int_{\theta} D^2 f(\theta \mid G = F) d\theta, \tag{2}$$

where $f(\theta|G = F)$ is the focal group density at θ. The DTF index is the population-weighted mean squared distance between the TCCs for the two groups. Moreover, DTF decomposes into differential item functioning at the item level:

$$\mathrm{DTF} = \sum_{i=1}^{n} \mathrm{CDIF}_i . \tag{3}$$

Violations of measurement equivalence between levels of *G* are evidenced by differential item and test functioning. DTF (as well as DIF) is identified when

$$f(Y \mid \theta, G = F) \neq f(Y \mid \theta, G = R). \tag{4}$$

That is, when individuals of the same ability θ belonging to different groups *F* and *R* do not have the same probability distribution for *Y*. A group difference remains even after controlling for θ by examining conditional probabilities at the same levels of ability.

ANALYSES

Analyses geared to document the internal structure of a test seek to gather information about the number of latent dimensions, the degree of linkage of test items and latent dimensions, and their generalization across relevant subpopulations. An internal structure analysis would document the degree to which the meaning and use of the test score can be recreated from the applicants' test performance on item-level data and whether the test is unbiased toward particular subpopulations of interest.

Item factor analysis was conducted separately on all PSU tests using MPlus (Muthén & Muthén, 2007). Note that for the Science tests, the analyses were performed on the common Science portion, as well as for each combination of the common portion with the specific portions (i.e., Physics, Biology, and Chemistry). The objective of these analyses was to determine whether each of the PSU tests overwhelmingly represents a single underlying factor or whether any of these tests show consistent evidence of a multi-factorial structure. This objective is addressed by comparing the sizes of the first three eigenvalues estimated from item level tetrachoric correlation matrices.

DTF was followed to evaluate the invariance of factor structure across subpopulations. The international evaluation team relied on a univariate IRT framework, after checking for unidimensional PSU solutions, for analyzing DTF and DIF. The absence of DTF is indicative that PSU scores are directly comparable on subpopulations, while the presence of DIF will show that item scores are inconsistent among subpopulations. Additionally, we present basic descriptive statistical information for these variables by showing the respective frequency counts for each response category on each form separately and for their total frequency across both forms of the tests.

Three parameter logistic model IRT item calibrations are used to attempt to fit a response function to each item by estimating a separate set of item parameter values for each item. The status of the items in relation to the latent factor is considered by examining how well the item response data fit the item response function.

This can be established by examining RMSE goodness of fit estimates, representing the average misalignment of item responses with the item response function, although the best and most informative assessment is often achieved by visually inspecting graphs showing the alignment of the item response data and item response functions.

Chi-square tests of goodness of fit, together with associated tests of statistical significance, have also been used to examine goodness of fit. However, with large samples of respondents—such as those for the PSU tests—even the slightest misalignment between the response data and response functions will produce chi-square values that are small in magnitude but statistically significant due to the large sample size. For this reason, statistical tests of significance do not provide a good criterion by which to judge goodness of fit.

In DIF analysis, IRT item parameters are obtained separately both for focal groups of substantive interest and for a reference group used for comparison purposes. To obtain these item parameters, item responses must be available in representative samples of observations from each of these groups. For these analyses, random samples were used to represent both the focal and reference group populations. The majority demographic group is typically used as the reference population. For this study, the following reference groups are used throughout:

- Gender: Male
- Region: Central
- Socio-economic Status: Average (Quintile C)
- Curricular Branch: Scientific-Humanistic
- Type of Financing: Municipal

Item response theory (IRT) provides a unified framework for investigating DIF (Raju, et. al., 1995). IRT is designed to represent what happens when an examinee encounters an item

on a test. The primary resource in IRT is the item response function, which models the probability of a correct response at different levels of ability. One such function for a dichotomous item is the IRT three-parameter logistic (3-PL) model, which includes an item difficulty, discrimination and guessing parameters.

A response function for dichotomous items represents the probability $P_i(\theta)$ that an applicant with ability θ will successfully complete item i. From a somewhat different perspective, this function also represents the proportion of applicants who will successfully complete this item at each level of ability. The response probability is represented by an s-shaped curve that rises monotonically with ability over the ability range $\{-\infty < \theta < \infty\}$.

PARTICIPANTS

A total of 231,140 examinees took the PSU for university admissions in 2012. All of the 231,140 students took both the Language and Communication test as well as the Mathematics test. The other tests were optional, with 140,114 examinees also taking the History and Social Sciences test. There is only one item set for each of these tests, so for Language and Communication, Mathematics, and History and Social Sciences there are three item sets.

The Science test was offered as an elective, with three alternate versions of the test provided in the main subjects of Biology, Physics and Chemistry, respectively. Short sections of each of these tests contained 18 common questions to assess core knowledge in the remaining two Science areas. With one main subject section and two common item sections, there are three Science tests but six item sets. Each of the Science tests totals about 80 questions, including 46 items in the main Science subject and 18 items for each of the two common item sets. (Note: In the 2012 administration DEMRE dropped one item from scoring of the following forms: 151/152, 161/162, and 171/172.)

Altogether, 132,969 examinees took one of the alternate Science tests, with 75,953 taking Biology, 27,143 taking Physics and 29,873 examinees taking the Chemistry version of the Science elective test.

INSTRUMENTS

The six PSU tests were each presented using two alternate forms, which we shall refer to as Form 1 and Form 2 on each respective test to collectively distinguish odd numbered forms from even numbered forms. The content of each form was identical, except for the relative order of the items on each form. Unique identifiers were not included to identify the items on each form. This required us to develop short labels for the items on each respective form, and then to use these item labels to identify individual items on each of the two forms. Our item naming conventions are based on single-letter designations for each test (L, M, H, B, F or Q) plus the relative positions of the items on Form 1 of each test. Thus, on the Language and Communication test, the 78 items are identified on both Forms 1 and 2 by the series of item names L01—L78.

Reading the PSU data file inevitably involves some inferences on our part as to the identities of individual items on Forms 1 and 2 for the Language and Communication, Mathematics, History and Social Sciences, Science-Biology, Science-Physics and Science-Chemistry tests. Hence, we need to prove beyond a reasonable margin for doubt that identical items in different positions on the two forms are being input correctly.

RESULTS

In this section, findings from item factor analyses and IRT-differential test functioning are presented in the first two sections. Distributions of n-counts by variables are available in Appendix C of this report.

Item Factor Analyses

Analyses were conducted separately on all six PSU item sets using MPlus software (Muthén & Muthén, 2007). The first objective of these analyses is to determine whether each of the six PSU item sets overwhelmingly represents a single underlying factor or whether any of these item sets shows consistent evidence of a multi-factorial structure. This issue is addressed by examining the ratios between the first three eigenvalues, labeled F1, F2, and F3 in Table 114. Evidence of a single underlying dimension is obtained by comparing the ratio of the difference of eigenvalues F1 and F2 with the difference between eigenvalues F2 and F3. When this ratio (Divgi's Index) is greater than 3, there is strong evidence of a single overall factor rather than a multi-factorial structure (see Hattie, 1985).

Table 114: Dimensionality Analysis Outcomes

Test	F1	F2	F3	F1 - F2	F2 - F3	Divgi's Index
Language	21.22	2.12	1.52	19.10	0.60	31.88
Mathematics	35.70	4.48	1.38	31.22	3.10	10.07
History	31.35	2.12	1.59	29.23	0.53	54.84
Science-Biology	16.26	1.58	1.35	14.69	0.23	64.41
Science-Physics	17.49	1.87	1.38	15.62	0.49	31.68
Science-Chemistry	12.86	1.66	1.20	11.20	0.46	24.30
Science-Common	23.99	1.67	1.14	22.32	0.53	42.04

As shown in Table 114, for each test, the first eigenvalues are quite large relative to the second or third eigenvalues. The differences between the first and second eigenvalues are largest for Mathematics (31.22) and smallest for Chemistry (11.20). The differences between the second and third eigenvalues are largest for Mathematics (3.10) and smallest for Biology (0.23). In this table, Divgi's Index ranges from a low of 10.07 for Mathematics to a high of 64.41 for Biology. It is interesting to note that all three Science tests, as well as the common portion, also show high values for Divgi's Index. The fact that the value of Divgi's Index is greater than 3 for all of these tests provides evidence of a single overall factor for each of the six PSU tests and the Science-Common item set.

IRT Differential Test Functioning (DTF)

Three-parameter logistic (3-PL) IRT analyses were performed to calibrate PSU items (see Appendix D for IRT results). The three parameters respectively represent item difficulty, item discrimination and item guessing. These three item parameters and their relationship to the response function are illustrated in Figure 38, using one of the items from the PSU Language and Communication test as an example.

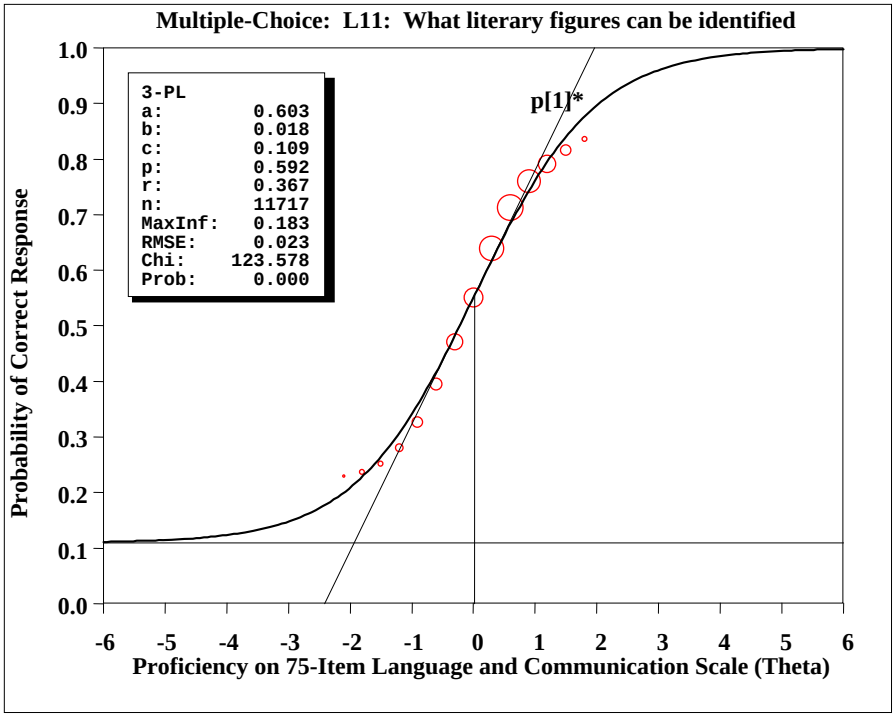


Figure 38: Item Characteristic Curve and Item Response Data for L11- *What literary figures can be identified* on the 75-Item Language and Communication Scale. Binary 3-PL L11 Difficulty Level is Appropriate b = 0.018, Power of Discrimination is Adequate a = 0.603, and Guessing is Low c = 0.109.

The latent proficiency distribution N(0, 1) runs across the x-axis of the figure and has a standard normal N(0, 1) frequency distribution, with mean $\mu = 0$ and standard deviation $\sigma = 1$. The latent proficiency distribution sets the scale for IRT proficiency estimates θ and item difficulty parameters b so that persons and items are positioned opposite one another along the x-axis, as illustrated by item difficulty parameter b in the figure.

The response function represents the probability of a correct response to the item at different levels of ability, while the IRT item parameters alter the shape and position of the response function. As shown by the response function in the figure, as person proficiency increases along the x-axis, probabilities rise as reported along the y-axis. As proficiency increases linearly, the probability of a correct response increases in a curvilinear fashion, following the s-shaped logistic curve in the figure.

The item difficulty parameter b = 0.018 is represented in the figure by the solid vertical line near the median of the latent proficiency distribution at zero and rising up to the response function. Higher values of difficulty parameter b represent progressively more difficult items and dislocate the response function to the right of the figure, without altering the response

function's slope or lower asymptote. Lower values of b move the line in the opposite direction and represent progressively easier items.

The item discrimination parameter a = 0.603 is represented by a solid tangent line intersecting the response function and rising from the lower left to upper right of the figure. Increasing values of a represent more discriminating items and will increase the slope of the tangent line so that it is more nearly vertical. Lower values of a represent lower discrimination and produce a tangent line and response function that rises more slowly from left to right in the figure.

The item guessing parameter c = 0.109 is represented by a solid flat line running across the figure at the lower asymptote of the response function found at the lower left of the figure. Increasing values of c represent higher levels of guessing by low proficiency examinees and will raise the asymptote of the function to a higher level of probability. Lower levels of c represent lower levels of guessing and will lower the asymptote to a lower level of probability. This takes guessing into account when tests are scored.

Notice that Language and Communication item L11 shown in the figure shows some evidence of item misfit, with chi-square = 125.178 and an associated p-value of 0.000. On the other hand, the typical error around the response function has a root mean square value of only RMSE = 0.025, or only an average 2.5 percentage points difference in relation to the response function. On the basis of the RMSE value and our own visual inspection of the relationship between the item responses—represented by circles scaled to size to represent population concentration—and the response function shown in the figure, we would probably conclude that this item has an acceptable degree of fit, even though this is less than perfect.

As shown in the caption to the figure, "Binary 3-PL L11 difficulty level is appropriate b = 0.02, power of discrimination is adequate a = 0.60, and guessing is low c = 0.11." classical test theory (CTT) difficulty is provided by the overall proportion of correct responses p = 0.592—including some random guessing—and classical discrimination, represented by the Pearson item to raw score correlation r = 0.367. All things considered, we would probably give this item a clean bill of health, despite the elevated chi-square value.

There are also four items with weak discrimination that essentially provide no useful information for discriminating levels of examinee proficiency. These items include:

- Science – Biology item
    - B37—"En una planta de tabaco se inocularon dos cepas…,"
- Language and Communication items
    - L01—"El tipo de mundo literario…,"
    - L30—"El condombé," and
    - L80—"Este fragmento corresponde a un …."

While these items could be excluded from the PSU test without loss of information, precision or reliability, their role in IRT scaling and scoring is innocuous except for lack of item fit.

The IRT framework for DIF examines between-group differences in the item response function representing a difference in the conditional probability of a correct response f(Y|θ), as shown in Figure 39. The figure shows the population-weighted vertical squared distance between the two response functions. The square root of this value $\sqrt{0.005}$ = 0.071, or about 7.1 percentage points, represents the average vertical distance between the two curves. Most of this difference is found toward the left of the figure, where differential rates of guessing are evident.

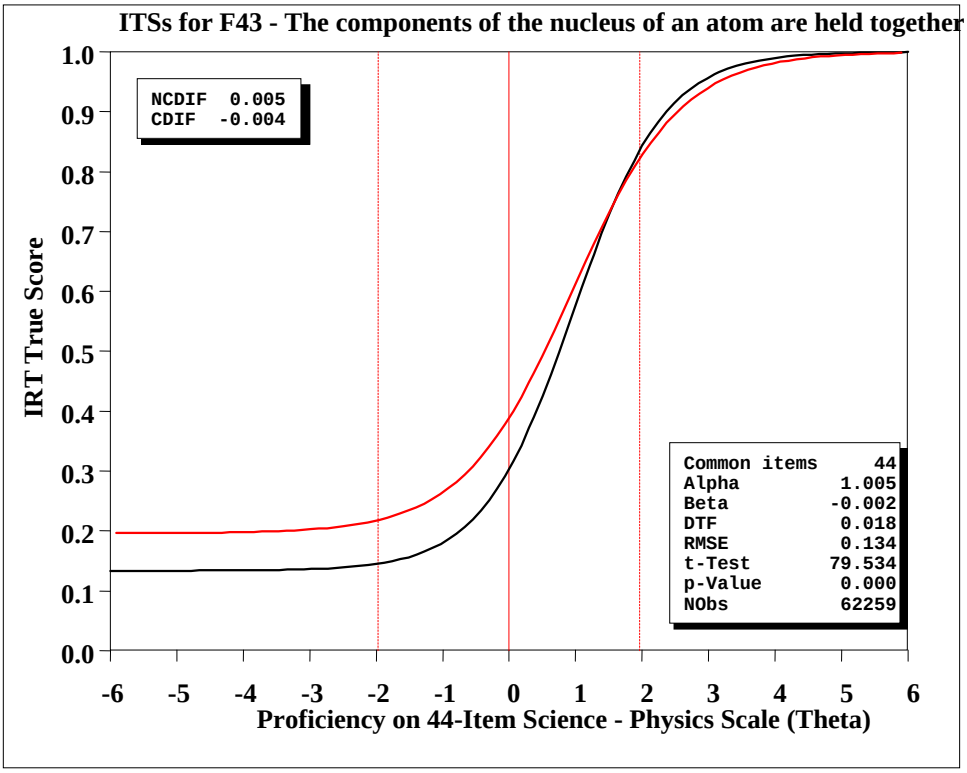


Figure 39: A Comparison of Response Functions for Two Groups on a 3-PL PPVT Item on the PSU Science–Physics Test

By contrast, *CDIF* is a compensatory DIF index that takes into account the compensating bias across all items:

$$CDIF = Cov(d_i D) + u_{d_i} u_D,$$

based on the covariance between $d_i = P_{iF}(\theta) - P_{iR}(\theta)$, the difference between the two response functions, and *D* = TF - TR, the difference between the expected proportion correct on the two tests, and the product between u_{d_i} and u_D, which are the respective means of d_i and *D*. For example, *CDIF* = -0.004 shows that this item is responsible for appreciable compensating bias favoring the focal group. At low and extremely high levels of ability (where there are few if any applicants), the focal group is more likely to answer this item correctly. Differential rates of guessing are apparent at low levels of ability.

PSU tests contain samples of items, where the DIF tendencies of individual items will often cancel one another out. Moreover, administrative decisions are always based on test results

and scarcely ever on the outcome for an individual item. For these reasons, it is often more informative to consider how the individual item-level DIF indices aggregate across all *n* items to affect the final test score. This examines differential test functioning (DTF) rather than the DIF for individual items. DTF will show whether an item sample will have a substantial impact on test scores in the two groups when the item responses are considered in the aggregate. Where total test scores are used to assess proficiency, DTF is likely to be more relevant than item-level DIF.

Table 115 reports test reliability, mean standard errors and the number of items for the six main PSU tests, excluding the common items on the Science elective tests. This shows that the tests are in general reliable—particularly on the longer tests—and therefore suitable for making decisions for individuals. The standard errors have correspondingly short intervals, in the vicinity of 0.19-0.32σ. However, to make informed decision, additional information is needed, such as the conditional standard of errors of measurement and indices of classification consistency.

Table 115: Test Reliability, Mean Standard Error of Measurement and Number of Items for the Six Main PSU Tests

Test	N Items	Reliability	SEM
Language and Communication	78	0.918	0.266
Mathematics	74	0.955	0.192
History and Social Sciences	75	0.945	0.214
Science – Biology	43	0.889	0.315
Science – Physics	44	0.903	0.291
Science – Chemistry	44	0.902	0.293

Everything depends on the intended use of these tests and where, within the latent proficiency distributions, priority decisions need to be made. As a practical matter, item difficulty and discrimination should align with the locations on the proficiency scale where important decisions are made. Objective 1.1.i provides supplemental information about CSEM for PSU tests.

Invariance across Subpopulations

In this section we address the invariance of factor structure for individual tests by major subpopulations. IRT provides a unified framework for examining DTF. For a test containing i = 1 ... n dichotomously scored items, the IRT true-score (T) is equal to the expected proportion correct on the test:

$$T = \sum_{i=1}^{n} P(Y_i = 1 \mid \theta).$$

The probabilities of correct response $P(Y_i = 1|\theta)$, when summed over the n items in a test, yield a model-based estimate of the number-right raw score, known as an IRT true-score or a test characteristic curve (TCC). Like the item response functions on which it is based, T also has an s-shaped curve that rises together with ability.

DTF procedures are illustrated for the PSU focal group (Technical-Professional) and PSU reference group (Scientific-Humanistic). The methods used in DTF analysis are demonstrated in a dramatic way by examining the performance of the Technical-Professional focal group on the Mathematics test, showing how the Stocking & Lord (1983) true-score equating procedure is used to place the separate item calibrations on the same

scale in order to obtain a common stratifying variable θ. This example is especially useful for conveying the logic of DTF analysis.

Item calibrations are obtained separately for each group. IRT sets the scale metric in normal deviates so that each group will initially have an N(0, 1) ability distribution, with latent ability mean $\theta = 0$ and standard deviation $\sigma_\theta = 1$, as shown for both the focal and reference groups in Figure 41. Ability estimates are expressed as normal deviates representing relative positions of examinees in each respective latent ability distribution. Ability estimates in the two groups are not directly comparable prior to equating, where the N(0, 1) metric represents relative positions separately within each group rather than directly comparable levels of ability.

This is easily verified by examining the relative positions of the two test characteristic curves shown in Figure 40. Raw scores in black for the reference group rise sharply in the lower tail of the ability distribution, beginning two standard deviations below the median, at -2σ. The curve has already risen appreciably to 40 percent by the time it reaches the middle of the reference population distribution at 0σ. The PSU Mathematics test is appropriately targeted for the majority reference group and only slightly too difficult. On the other hand, for the focal group, raw scores on the same test only begin to rise above -1σ, well into the focal group distribution, near the center of Figure 40, at the focal group's 16th percentile. The red line positioned far to the right of the black line shows that the PSU Mathematics test is much more difficult for Technical-Professional students. One has to go to a higher percentile in the Technical-Professional distribution before finding someone with the same number-right score found in the reference population.

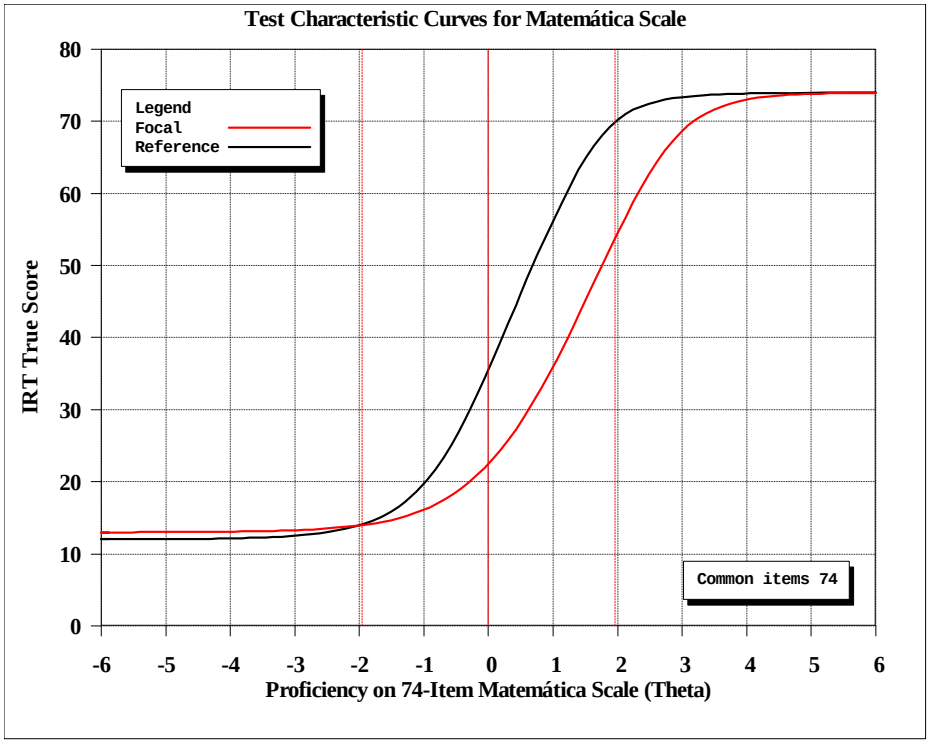


Figure 40: Focal and Reference TCCs *before* Equating

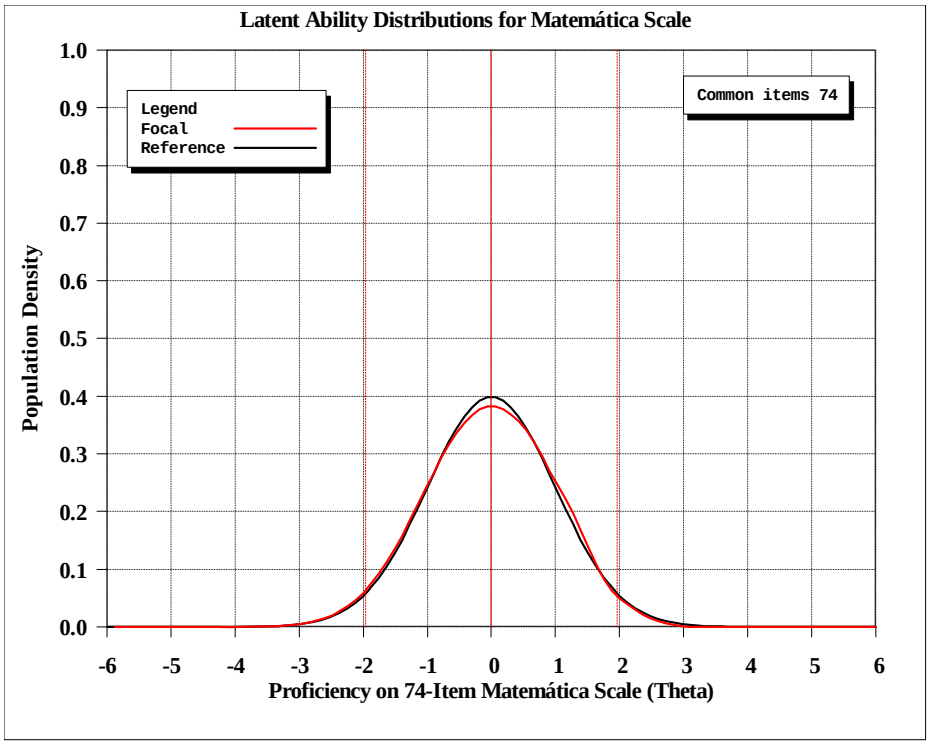


Figure 41: Focal and Reference Latent Ability Distributions *before* Equating

To properly reflect the same absolute levels of ability in both groups, we need to align the number-right scores in the two groups. Since the reference group is the standard of comparison, the red curve of the focal group on the right needs to move left so that it aligns as closely as possible with the black curve of the reference group. To place focal and reference test results on the same scale metric, IRT true-score equating is used to find the simple linear transformation of slope α and of origin β that provides the closest alignment in terms of the model-based number-right raw score *T*. When model-based raw scores align as closely as possible across the full range of ability, the tests are said to have been equated. The tests have been equated because students who obtain the same scores on the same items receive the same IRT ability estimates. After equating, IRT ability estimates are reported on the same scale metric and are therefore directly comparable. DTF and DIF analyses require that the two tests be as closely aligned as possible before a DTF misfit index is calculated.

To obtain comparable results, IRT true-score equating is applied using the common items in both tests. True-score equating finds the simple linear transformation of origin β and of scale α that will align the TCCs as closely as possible across the entire ability distribution, as shown in Figure 42. A linear transformation of slope α = 0.801 and of origin β = -0.761 aligns the TCCs for the PSU Mathematics test as closely as possible in both of the two groups. Both coefficients are reported in reference population standard deviation units (σ). This implies that the Technical-Professional ability estimates should be multiplied by α = 0.801 before β = -0.761 is added to the result to obtain ability estimates reported in the reference population scale metric. Technical-Professional students have much lower scores on the English-language screener relative to the reference group, and the variation in ability is much more restrictive.

Equating constants α and β describe how the focal group population is related to the reference group population. Once the two sets of tests have been equated using their respective test information functions (TIFs), the Raju, et. al. (1995) fully parametric differential test functioning (DTF) and differential item function (DIF) analyses are run to assess test and item consistencies on the two sets of forms.

The test characteristic curves presented in Figure 42 purport to show how closely the two test characteristic curves align after equating. However, the scale of the graph is such that two TCCs appear to be almost perfectly superimposed. We can see that the two curves are not perfectly superimposed by inspecting the DTF index and RMSE reported in the box at the right of Figure 42, but it is not easy to actually see any difference between the two curves due to the scale of the graph. With an average difference between the test TCCs, the unsigned RMSE = 0.484 coefficient shows that on average the two curves differ by nearly half a raw score point across all ability levels and shows that there is indeed some differential test functioning evident in these two groups, as seen along the vertical axis of Figure 42. To keep this number in perspective, Scientific-Humanistic students will average 40 percent of 74 or 30 raw score points on the Mathematics test. Differential test functioning will thus contribute errors of 0.484 / 29.6 = 0.016 or between 1 and 2 percentage points on this PSU test.

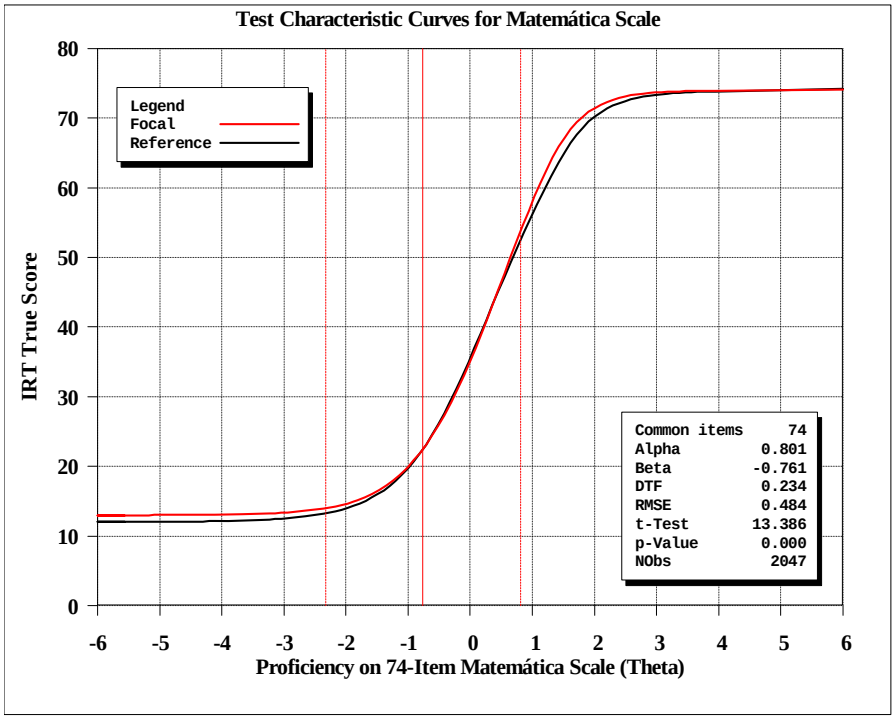


Figure 42: Focal and Reference TCCs *after* Equating

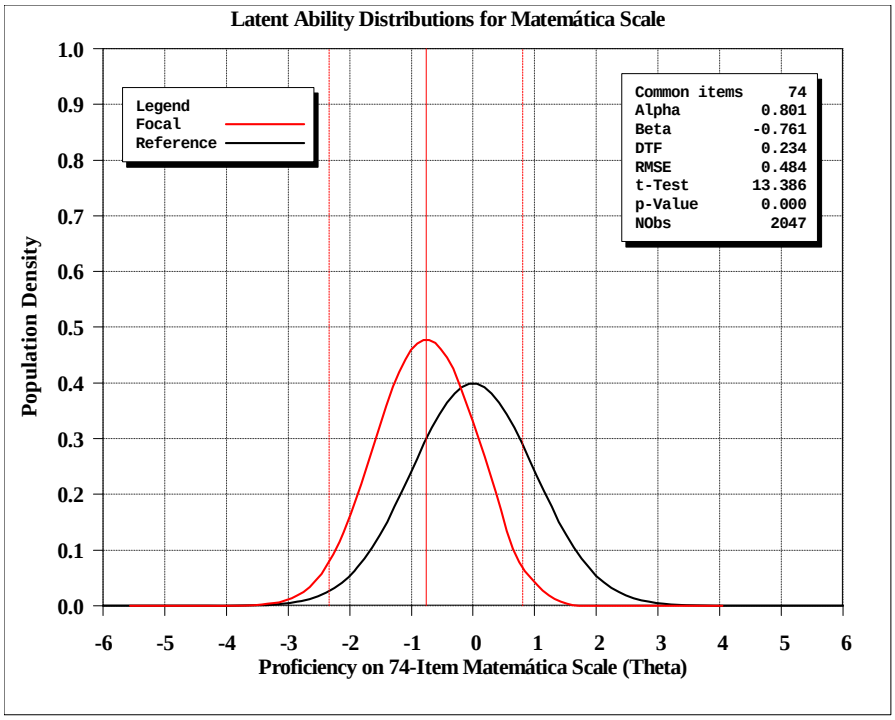


Figure 43: Focal and Reference Latent Ability Distributions *after* Equating

Table 116 reports focal and reference group comparisons for selected demographic groups on all six PSU tests. Equating constants alpha (α) and beta (β), representing respectively the change in dispersion and location, show how the focal latent ability distribution relates to the corresponding reference group population. The DTF index represents the squared differences between the two respective test characteristic curves after equating, while the root mean squared value RMSE is the square root of the index value, reported in raw score points. In the final column of the table, we have used the criteria of 0.020 raw score point to represent substantively meaningful differential test functioning that may warrant further consideration.

Generally speaking, PSU tests show little evidence of differential test functioning. The strongest evidence of DTF is seen for the subpopulations based on SES, curricular branch, and type of school. For the SES, the Mathematics test and the Language and Communication test showed DFT for most comparisons. Differences favored consistently the higher SES in the pair-comparison.

Regarding curricular branch, differences between lower performing Technical-Professional students relative to higher performing Scientific-Humanistic students was observed for Science–Biology, Biology (Common), Science-Chemistry, Language and Communication and Mathematics tests. Differences favor the latter group (Scientific-Humanistic group).
For type of school, differences between Private and Municipal, favoring the Private group, were found for Mathematics, Language and Communication, History and Social Sciences, and Science-Biology. For the comparison between Subsidized and Municipal schools DTF differences were found for Mathematics. The difference favored the Subsidized group.

Table 116: Differential Test Functioning on PSU Tests for Selected Subpopulations

Focus	Reference	Test	Alpha Slope *Sigma σ*	Beta Intercept *Mu μ*	DTF Index	DTF RMSE *Raw Score Point*	DTF evidence *Yes = RMSE > 0.20*
Female	Male	Language and Communication	1.028	-0.007	0.002	0.040	No
		Mathematics	0.979	-0.243	0.028	0.168	No
		History and Social Sciences	1.003	-0.251	0.007	0.083	No
		Science - Biology	1.023	-0.230	0.023	0.152	No
		Science - Physics	0.954	-0.154	0.014	0.119	No
		Science - Chemistry	0.887	-0.170	0.020	0.142	No
		Biology - Common	0.957	-0.098	0.000	0.029	No
		Physics - Common	0.940	-0.300	0.023	0.151	No
		Chemistry - Common	1.021	-0.185	0.004	0.067	No
North	Central	Language and Communication	0.966	-0.099	0.002	0.045	No
		Mathematics	0.909	0.002	0.061	0.247	Yes
		History and Social Sciences	0.952	-0.151	0.049	0.222	Yes
		Science - Biology	0.952	-0.141	0.004	0.067	No
		Science - Physics	0.922	-0.324	0.011	0.103	No
		Science - Chemistry	0.965	-0.132	0.040	0.199	No
		Biology - Common	0.961	-0.283	0.020	0.142	No
		Physics - Common	0.936	-0.136	0.006	0.076	No
		Chemistry - Common	0.893	-0.118	0.008	0.092	No
South	Central	Language and Communication	0.955	-0.110	0.023	0.150	No
		Mathematics	0.932	-0.058	0.041	0.203	Yes
		History and Social Sciences	0.918	-0.083	0.020	0.141	No
		Science - Biology	0.990	-0.231	0.009	0.096	No
		Science - Physics	0.955	-0.209	0.010	0.101	No
		Science - Chemistry	0.987	-0.112	0.018	0.134	No
		Biology - Common	0.946	-0.124	0.008	0.088	No
		Physics - Common	0.975	-0.180	0.006	0.075	No
		Chemistry - Common	0.974	-0.144	0.007	0.085	No
NSE Low	NSE Medium	Language and Communication	1.097	-0.502	0.068	0.260	Yes
		Mathematics	1.016	-0.476	0.021	0.145	No
		History and Social Sciences	1.009	-0.436	0.028	0.168	No
		Science - Biology	1.009	-0.413	0.006	0.076	No
		Science - Physics	1.039	-0.449	0.002	0.047	No

Focus	Reference	Test	Alpha Slope *Sigma σ*	Beta Intercept *Mu μ*	DTF Index	DTF RMSE *Raw Score Point*	DTF evidence *Yes = RMSE > 0.20*
		Science - Chemistry	1.160	-0.519	0.005	0.073	No
		Biology - Common	1.134	-0.524	0.000	0.030	No
		Physics - Common	1.096	-0.502	0.013	0.116	No
		Chemistry - Common	1.015	-0.430	0.002	0.047	No
NSE Medium Low	NSE Medium	Language and Communication	1.041	-0.201	0.004	0.065	No
		Mathematics	1.023	-0.192	0.049	0.220	Yes
		History and Social Sciences	0.987	-0.187	0.005	0.068	No
		Science - Biology	1.006	-0.192	0.003	0.055	No
		Science - Physics	0.991	-0.170	0.015	0.122	No
		Science - Chemistry	1.032	-0.184	0.003	0.059	No
		Biology - Common	0.998	-0.158	0.006	0.076	No
		Physics - Common	1.011	-0.266	0.017	0.129	No
		Chemistry - Common	1.010	-0.234	0.003	0.055	No
NSE Medium High	NSE Medium	Language and Communication	0.928	0.290	0.140	0.374	Yes
		Mathematics	0.953	0.332	0.195	0.441	Yes
		History and Social Sciences	0.941	0.214	0.022	0.149	No
		Science - Biology	0.953	0.282	0.002	0.049	No
		Science - Physics	0.970	0.215	0.002	0.044	No
		Science - Chemistry	0.966	0.206	0.002	0.048	No
		Biology - Common	0.939	0.242	0.005	0.071	No
		Physics - Common	0.983	0.231	0.023	0.151	No
		Chemistry - Common	0.950	0.216	0.002	0.040	No
NSE High	NSE Medium	Language and Communication	0.791	0.780	0.116	0.341	Yes
		Mathematics	0.858	0.956	0.237	0.487	Yes
		History and Social Sciences	0.890	0.777	0.003	0.057	No
		Science - Biology	0.938	0.879	0.008	0.090	No
		Science - Physics	0.933	0.816	0.009	0.095	No
		Science - Chemistry	0.943	0.736	0.008	0.092	No
		Biology - Common	0.868	0.747	0.002	0.045	No
		Physics - Common	0.913	0.785	0.005	0.073	No
		Chemistry - Common	0.919	0.805	0.001	0.035	No
Technical-Profession	Scientific-Humanistic	Language and Communication	0.918	-0.776	0.088	0.297	Yes
		Mathematics	0.847	-0.874	0.189	0.435	Yes
		History and Social Sciences	0.919	-0.681	0.009	0.094	No

Focus	Reference	Test	Alpha Slope *Sigma σ*	Beta Intercept *Mu μ*	DTF Index	DTF RMSE *Raw Score Point*	DTF evidence *Yes = RMSE > 0.20*
		Science - Biology	0.879	-0.769	0.113	0.337	Yes
		Science - Physics	0.840	-0.790	0.014	0.117	No
		Science - Chemistry	1.038	-0.914	0.111	0.333	Yes
		Biology - Common	0.937	-1.005	0.085	0.292	Yes
		Physics - Common	0.861	-0.805	0.005	0.073	No
		Chemistry - Common	0.912	-0.978	0.012	0.108	No
Private	Municipal	Language and Communication	0.717	1.100	0.129	0.359	Yes
		Mathematics	0.805	1.380	0.128	0.358	Yes
		History and Social Sciences	0.878	1.075	0.042	0.204	Yes
		Science - Biology	0.861	1.194	0.056	0.236	Yes
		Science - Physics	0.789	1.065	0.006	0.077	No
		Science - Chemistry	0.852	0.921	0.023	0.153	No
		Biology - Common	0.770	0.898	0.002	0.044	No
		Physics - Common	0.814	1.077	0.017	0.130	No
		Chemistry - Common	0.864	1.080	0.007	0.084	No
Subsidized	Municipal	Language and Communication	0.926	0.304	0.004	0.061	No
		Mathematics	0.925	0.380	0.115	0.339	Yes
		History and Social Sciences	0.960	0.245	0.020	0.140	No
		Science - Biology	0.966	0.280	0.031	0.176	No
		Science - Physics	0.864	0.264	0.015	0.124	No
		Science - Chemistry	0.908	0.166	0.007	0.081	No
		Biology - Common	0.929	0.176	0.001	0.038	No
		Physics - Common	0.920	0.287	0.004	0.064	No
		Chemistry - Common	0.912	0.248	0.009	0.097	No

EVALUATION

Disparities observed in mean scores on standardized tests by different sub-groups often give rise to allegations of bias in testing. Tests would be biased against minorities if they contained content from outside the realm of minority cultural experience by dealing with content that minority students are unfamiliar with or have had little opportunity to learn. The possibility that a test contains language and content to which minority groups have had limited exposure has created perceptions of unfairness in testing commonly referred to as bias. Since tests and testing practices have come under close public scrutiny, test publishers and practitioners routinely provide information about test bias.

Analyses geared to document internal structure of a test seek to gather information about the number of latent dimensions, the degree of linkage of test items and latent dimensions, and their generalization across relevant subpopulations. An internal structure analysis would document the degree to which the meaning and use of the test score can be recreated from

applicants' test performance on item level data and whether the test is unbiased toward particular subpopulations of interest.

The factor analytic framework provided means to investigate the link between item-level data and their corresponding underlying latent dimensions. Findings from the analyses supported the presence of a strong latent dimension for PSU tests. Such a finding is encouraging and supports future use of one-dimensional item response theory models to set and maintain PSU scales over years of test administration and the equating of forms within a given administration. In Chile, PSU test scores from a given administration are valid for two years; thus, comparability of test scores is necessary for building a fair national admission testing program. The existence of a "single underlying dimension" for each of the tests is a necessary but insufficient condition for test validity. In our analysis the evaluation team found a single underlying dimension for each of the PSU tests (Language and Communication, Mathematics, History and Social Sciences, Science–Common, Science-Biology, Science-Physics, and Science-Chemistry. This finding does not necessarily mean that the underlying dimension for each test is one and the same, even for the various Science tests.

The *Standards for Educational and Psychological Measurement* (1999) describes fairness as a correspondence between equal probabilities of success for groups of test takers of similar standing on an ability continuum. Under the item response theory framework, differential test functioning provides a means to investigate the lack of invariance of links between observed item responses and their latent underlying variable. Generally speaking, PSU tests show little evidence of differential test functioning (DTF). In these circumstances, it is reasonable to conclude that factor structure invariance of tests by subpopulation groups has been achieved. Particularly speaking the strongest evidence of DTF is seen for lower performing Technical-Professional students relative to higher performing Scientific-Humanistic students on the Science – Biology, Language and Communication and Mathematics tests and for much higher performing Private and somewhat higher performing Subsidized students relative to less well-performing Municipal students on the Language and Communication and Mathematics tests. These are specific cases that may warrant further consideration and review.

With respect to the results by PSU test subject, Mathematics showed the largest number of DTF flags with eight flags out of the ten sub-group comparisons. This was followed by Language and Communication with five DTF flags.

It is important to consider where within the latent proficiency distributions priority decisions need to be made. As a practical matter, item difficulty and discrimination should align with the locations on the proficiency scale where standard errors are short and important decisions are made. Thus, it is not so much a question of overall reliability, but rather precision at the likely points of decision-making that should drive the design of university admissions tests. Our analyses have shown that there are four items with weak discrimination that essentially provide no useful information for discriminating levels of examinee proficiency. As noted above, these items include:

- Science – Biology item
    - B37—"En una planta de tabaco se inocularon dos cepas…,"
- Language and Communication items
    - L01—"El tipo de mundo literario…,"
    - L30—"El condombé," and
    - L80—"Este fragmento corresponde a un …."

While these items could be excluded from the PSU test without loss of information, precision or reliability, their role in IRT scaling and scoring is innocuous except for lack of item fit.

RECOMMENDATIONS

1. We recommend adopting the IRT framework for test construction activities, item-level analyses, scaling, and scale maintenance.
2. We recommend the careful selection of operational items during test construction activities so that high levels of precision at the critical decision point of the score scale are obtained.
3. We recommend using the factor analysis results showing the unidimensionality of the PSU to ground the use of IRT to scale and equate the PSU.
4. As a result of the evaluation team's differential test functioning (DTF) analyses, it recommends that the PSU program conduct additional analyses to understand better the DTF between private and subsidized schools versus municipal schools, particularly for the Language and Communication and Mathematics tests. This is a recognized standard (Standard 7.3, AERA, APA, & NCME, 1999) for high-stakes test development worldwide where the fairness of the test across different subpopulations is an issue.

BIBLIOGRAPHY

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Arce-Ferrer, A., & Corral-Verdugo, V. (2002). La medición de la aptitud académica general: Estudio de un caso en el ingreso a la licenciatura. *Revista Mexicana de Psicología, 19*, 1, 57-72.

Baker, F., & Kim, S. (2004). *Item response theory. Parameter estimation techniques.* Second edition. New York: Marcel Dekker, Inc.

Bock, R., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261-280.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

du Toit, M. (ed.) (2003). IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT. Lincolnwood, IL: Scientific Software International, Inc.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Press.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Joreskog, K. (1994). On estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381-389.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* 4th edition. (pp. 17-64). Westport: American Council on Education and Praeger Publishers.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores, with contributions by Alan Birnbaum*. Reading, MA: Addison-Wesley.

McDonald, R. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (1998-2007). *MPlus user's guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *v.19*, n.4 (December), p. 353-368.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Objective 2.2. Content validity: Logical and empirical analyses to determine if the test content fully reflects the domain and if the test has the same relevance regarding the interpretation of the scores in the population subgroups

ABSTRACT

This document reports detailed results of Phases 1 and 2 of the content validity study of the PSU. The core component of Phase 1 is an alignment study geared toward investigating degree of coverage of the intended domain of the PSU tests administered for the 2012 admissions process. This effort took into consideration Chile's national curriculum, OF and CMO for the Scientific-Humanistic and Technical-Professional curricular branches. The analyses were performed considering General Training (four years of high school) and Differentiated Training (Scientific-Humanistic and Technical-Professional).

Students taking the PSU are required to take the "Language and communication" and "Mathematics" exams, and then must take either the "History and Social Science" exam or one of the three natural Science exams (Biology, Chemistry, or Physics). In this study, all of the PSU exams were aligned to curriculum standards (CMO and OF) used by the Chilean Ministry of Education. The analyses following the methodology of Webb (1997) included four alignment criteria:

(1) categorical concurrence, which addresses whether the assessment covers the same broad content categories,
(2) depth-of-knowledge consistency, which compares the level of cognitive complexity required by the standards with those required by the test items,
(3) range-of-knowledge correspondence, which examines the breadth of knowledge assessed compared to the expectations set forth in the standards, and
(4) balance of representation, which examines the distribution of assessment items across content objectives.

This report contains five sections. The two appendices are provided at the end of the entire Evaluation of the PSU. Section 1 introduces the study topic and brings high level description of PSU tests and Chile's National Curriculum. Section 2 describes the Webb alignment method in general and the approach employed to carry out the assessment of the PSU informing Phase 1. In addition, the relevant qualifications of content specialists involved in the study are summarized. Section 3 contains the results of the alignment of the PSU to the curriculum framework as well as pertinent standard sets (CMO) from both the Scientific-Humanistic and Technical-Professional curricular branches. Section 4 includes the summaries of interviews with key stakeholders that form the Phase 2 analysis as well as the interview protocols. It also includes a statement from the Curriculum Unit of MINEDUC concerning the alignment between the PSU and the National Curriculum. Section 5 provides an overall evaluation base on the Phase 1 and Phase 2 analyses, recommendations for improving the content validity of the PSU tests and the bibliography. Appendix E deals with "sources of challenge," which lists item numbers of questions on the PSU assessments that were found to be problematic and an explanation of why they are problematic.

INTRODUCTION

The introductory section is organized around three major parts. The first one describes important decisions made when conceptualizing the depth and breadth of PSU content validity study. The second part describes in broad terms the policy context concerning the PSU. In the third part, a general overview of the PSU assessment is presented to familiarize readers with essential information on the college selection test. Finally, the fourth part is devoted to provide a high-level summary of Chile's National Curriculum.

Conceptualizing the Content Validity Study

The scope of alignment study initially described in the Pearson proposal has undergone fine tuning. Regarding content validity, the evaluation team added the following to the study:

- When evaluating PSU content validity, analyses involving Technical-Professional branch of Chile's National High School Curriculum should be added to those initially devised for Scientific-Humanistic branch of the curriculum. As part of the goal clarification meeting, Pearson refined the need for adding these analyses due to its relevance for interpreting PSU test scores for the subgroups (e.g., Scientific-Humanistic and Technical-Professional).

- When evaluating PSU content validity, analyses of both Fundamental Objectives (OF) and Mandated Minimal Contents (CMO) stated on 2009 Chile's high school national curriculum should be added to capture better the domain when analyzing the alignment of PSU test content.

During a March 2012 meeting on the nature of the evaluation of the objective, there was a request to add supplemental information on PSU content validity from an alternative domain to the declared domain declared in pieces of resolutions from *Consejo de Rectores de Universidades Chilenas* (CRUCH). Concerns about this approach were raised that such departure from policy supported domain (intended domain) would contradict policy for defining the domain of the PSU as stated by CRUCH and the *Comisión Nuevo Currículum de la Enseñanza Media y Pruebas del Sistema de Admisión a la Educación Superior*.

In an effort to respond to these concerns, Pearson has added a layer of supplemental analyses to the initial content validity alignment study. This additional effort resulted in a set of refocused meetings with stakeholders (high school teachers and university teachers) in which they were asked for their perceptions of prospective students' academic readiness for university studies.

Policy Context

Pearson consulted relevant policy documentation defining the intended domain of the PSU battery and intended purpose of university admissions process to provide a framework to reference the analyses of content validity. The documentation was part of the literature review update Pearson carried out after the goal clarification and requirement collection meeting. The document was produced to fine tune the evaluation processes that defined the intended PSU domain when analyzing the meaning of PSU test scores. That is, the framework has guided development of evaluation activities, particularly those for content validity of PSU battery informing the 2012 university selection process. The following policy statements form part of the adopted framework:

> The battery of tests that make up the University Selection Test (PSU®) arises by order of the Council of Rectors of Chilean Universities (CRUCH) in 2001 and is based on the Fundamental Objectives (OF) and minimum

> compulsory contents (CMO) of the Secondary Education curriculum developed by the Ministry of Education (MINEDUC) in 1998. (MINEDUC, 1998)

A second aspect of the policy was identified as part of the review and conceptualization of Chile's selection process for universities belonging to the CRUCH. This aspect of the policy states:

> The purpose of the admissions process is to select the candidates that apply to be accepted in one of the twenty five institutions forming part of the CRUCH. The objective of the system is to select those applicants that obtain the best performances in the battery of tests composing the PSU, under the assumption that they represent the best possibilities of successfully complying with the tasks demanded by higher education, for them to enter according to their preference, to one of the institutions forming CRUCH, in the careers they are applying for. Said purpose is achieved by means of the application of educational measurement instruments (PSU), along with the inclusion of the average scores during Secondary Education (NEM).

Description of the PSU Assessment

The PSU is the national university admission test and has been required for entrance into many Chilean institutes of higher education since 2003. Currently, the PSU is required by the twenty-five Chilean universities belonging to the Consejo de Rectores de las Universidades Chilenas (CRUCH) as well as eight other private universities.

The PSU consists of two mandatory tests, *Language and Communication* and *Mathematics*, and four other tests, *History and Social Sciences*, *Biology*, *Physics* and *Chemistry*. For the latter group of PSU tests, students aspiring to attend a Chilean university must choose, based on their chosen course of university studies, between "History and Social Science" or "Science" (either Physics, Chemistry, or Biology). The PSU tests are administered on two consecutive days every December to all applicants registered to take them.

Description of Chile's National High School Curriculum

Chile's high school curriculum covers four years of *Formación General* (or General Training) for all students attending high school. Through their high school experience, the curriculum has a building in component to offer for possibilities of increasing coverage and depth of general formation contents during the last two years of high school. These latter characteristics of Chile's high school curriculum are known collectively as *Formación Diferenciada* or Differentiated Training. In Chile there are two kinds of Differentiated Training: (1) Scientific-Humanistic and (2) Technical-Professional. Students in the Scientific-Humanistic track choose a sub-area of focus (Science or the humanities) and attend classes offered at their schools buildings to prepare them for future university-level studies in these areas. In contrast, students in the Technical-Professional track choose a sub-area of focus (e.g., Chemistry) and learn competencies that will prepare them to enter the workforce in that field following graduation and continue learning through their lives.

METHODOLOGY

The PSU content validity study involves a two-phase process for evaluating relevant facets of the PSU test used for the 2012 admissions process. Phase 1 covers an alignment study to investigate the degree of coverage of the intended domain of the PSU tests administered for the 2012 admissions process. This effort took into consideration Chile's national curriculum, the OF and the CMO for Scientific-Humanistic and Technical-Professional branches. Phase 2 involves analyses of supplemental information on OF- and CMO-implemented curriculum in high school classrooms, and the OF and CMO relevant for students' readiness to begin college learning. Whereas Phase 1 uses an alignment study, Phase 2 relies on interviews with relevant stakeholders (i.e., high school teachers and university teachers).

Phase 1: Alignment Study

Pearson proposed the following the five dimensions of Webb's (1997) alignment process to judge the alignment between Chile's high school curricular content standards and the PSU tests.

- **Categorical Concurrence:** Does the PSU measure what the curricular standards state students should both know and be able to do?
- **Depth of Knowledge:** Does the PSU reflect the cognitive demand and depth of the curricular standards? Is the PSU as cognitively demanding as the standards?
- **Range of Knowledge:** Does the PSU reflect the breadth of the curricular standards?
- **Balance of Representation**: Does the PSU reflect the full range of the curricular standards?
- **Source of Challenge:** Does the PSU reflect cognitive demands extraneous to those in the curricular standards?

A typical Webb study involves a criterion for evaluating the degree of alignment between content standards and assessments. The following criteria have been used to develop summative statements of alignment.

- **Categorical Concurrence:** At least 6 items measures content for each standard.
- **Depth of Knowledge Consistency:** At least 50% of assessment items are as cognitively demanding as expectations stated on the standards.
- **Range of Knowledge:** At least 50% of the objectives per standard have one related assessment item.
- **Balance of Representation:** An index value of at least 0.7 is obtained based on the difference in the proportion of content benchmark and the proportion of items corresponding to the content benchmark (i.e., 70% of the objectives per standard have given equal amount of emphasis on the assessment).
- **Source of Challenge:** Less than 2% of the assessment items inadvertently bring cognitive demands other than those targeted by the objective (e.g., extraneous constructs). This element will be reported but not used to evaluate the strength of the alignment of the PSU to Chile's high school content standards.

Findings for each content area are reported for alignment criteria mentioned above with sources of challenge issues and other notes.

Table 117 shows Webb's classification system for the levels of alignment between assessments and content standards. This typology is used to report PSU degree of alignment to intended domain.

Table 117: Webb Alignment Level

Alignment Levels Using Four Criteria				
Alignment Level	**Categorical Concurrence**	**Depth of Knowledge**	**Range of Knowledge**	**Balance of Representation**
High	*6 items per standard*	*50%*	*50%*	*70%*
Medium	*---*	*40% - 49%*	*40% - 49%*	*60% - 69%*
Low	*Fewer than 6 items per standard*	*Less than 40%*	*Less than 40%*	*Less than 60%*

The Webb Alignment Method

The Webb method (Webb, 2005) was designed as an approach to assess 1) the degree to which standards (CMO and OF) are addressed by test questions, 2) the level of cognitive complexity required by test questions, 3) the breadth of knowledge required by the test questions, and 4) the evenness of coverage of standards (CMO and OF) by a test; thus giving a more comprehensive view of a test's potential to evaluate students on required material than would be possible by merely ensuring that all standards are addressed as part of an exam. Webb (1997) named the four alignment analyses mentioned above "categorical concurrence," "depth-of-knowledge consistency," "range-of-knowledge correspondence," and "balance of representation," respectively. Below, we will give brief definitions of each of these analyses, including how they are calculated as well as target levels for each type of analysis.

Categorical concurrence addresses the extent to which the assessment and the content strands (CMO and OF) cover the same content categories. For example, if the PSU math test addresses all of the Chilean Mathematics standards, then categorical concurrence will be high. Conversely, lower categorical concurrence means that many standards (CMO and OF) are left unaddressed by the exam. The target criterion for acceptable categorical concurrence is six test items per standard.

Depth-of-knowledge consistency evaluates the extent to which the assessment and the content expectations (CMO and OF) require the same level of cognitive complexity. This is done by assigning each test item and each standard a depth-of-knowledge (DOK) level according to a four-point scale (Table 118).

Table 118: Depth of Knowledge (DOK) levels (Adapted from Hess, 2005)

DOK level	Type of thinking	Explanation
1	Recall	Requires recall or recognition of facts
2	Basic application	Requires use of knowledge to do routine problems; organization of data
3	Strategic thinking	Requires reasoning, decision-making, and justification
4	Extended thinking	Requires research, inter-disciplinary connections, creativity

Next, the DOK level of the test questions is compared with that of the corresponding standards. Here, adequate consistency requires at least 50 percent of items to be at the same or higher DOK level than the strand they are designed to assess.

Range-of-knowledge correspondence explores the breadth of knowledge required for the standards and the assessment. To evaluate range-of-knowledge correspondence within each strand, it is necessary to examine the number of objectives that correspond with at least one test item. For the PSU, at least 50% of the objectives per standard should be aligned with at least one test item.

Balance of representation evaluates the distribution of assessment tasks across content objectives with the aim of ensuring that no standards within a content strand are either over- or under-represented. Balance of representation is calculated according to Webb (2005). Briefly, an index value is obtained based on the difference between the proportion of content benchmark and the proportion of items corresponding to that benchmark.

Finally, "source of challenge" is an additional measure of a test's alignment to standards that will be assessed. The purpose of this analysis is to ensure that the "sources of challenge" faced by a student on an assessment stem primarily from assessment items that are related to the standards (CMO and OF) and not from extraneous topics. Ideally, no more than 2% of the items in the PSU should present "sources of challenge." Results from this analysis are found in Appendix E.

Expert Panelists

The alignment study was performed with a Pearson team of content area specialists, all of whom are fluent in Spanish. Of the six panelists, two have traveled extensively or lived in Chile. Four of the panelists are teachers with an average of 8 years working in the classroom or with curriculum development. Four of the panelists have earned advanced degrees (masters or doctoral) in their subject areas.

Alignment Protocol

The Webb alignment of the PSU was performed as follows. First, the project team leader assigned the different portions of the PSU (Mathematics, Science, etc.) to the content area specialists who had been extensively trained in the Webb alignment method. The content area specialists' first task was then to assign depth-of-knowledge (DOK) ratings to the objectives of the Chilean Ministry of Education's high school content standards. They then similarly assigned DOK ratings to PSU assessment items. The panelists then determined the categorical concurrence, DOK consistency, range-of-knowledge correspondence, balance of

representation, and source of challenge criteria as defined above. Once completed, the content area specialists sent their work to the project team leader, who then sent it out for independent review.

Phase 2: Analyses of supplemental information on OF and CMO implemented curriculum, and OF and CMO relevant to begin college learning

Panels are in the process of being assembled with relevant participants identified by Chile's Ministry of Education and recruited by Pearson. When feasible, one day before each panel meeting, Pearson electronically distributes the goals of the meeting and the discussion protocols that will be used so that panel members can prepare.

The purposes of the meetings are to gain deeper understanding and documentation of stakeholders' anecdotal information:

- on the degree alignment of PSU domain to classroom instruction;
- that the PSU as a university selection test adequately indicates readiness of entry level university students at the beginning of college instruction;
- on relationship between Chile's national high school curriculum and level of knowledge required for entry level university students to be successful.

The evaluation team proposed using accessible groups of stakeholders to gather their anecdotal information on the PSU test domain and the relationship to levels of knowledge and skills relevant for entry level students to be successful. The following types of stakeholders were identified and recruited to participate in the meetings: high school teachers and university teachers. (Note characteristics of these groups were defined earlier in this report).

Interview meetings begin with Pearson delivering a high level introduction of the purpose and rules of the meeting and a general overview of the PSU evaluation. Following this presentation interviewees are asked to familiarize themselves with the OF and CMO on Chile's national high school curricular branches (Scientific-Humanistic and Technical-Professional) and to follow the directions stated in the protocol. The evaluation team facilitator encourages discussion and alternate points of view from the panel members.

We have requested that the meetings be recorded so that we can carefully document the statements provided by the panel members, which would otherwise be difficult to accurately record during the meeting. We will obtain permission from participants to record these meetings and adhere to all legal requirements for the maintenance of the recordings while in our possession.

Pearson proposed considering, when feasible, the following variables when locating and inviting accessible participants for meetings.

- Region (North, Central and South)
- Curricular branch (Scientific-Humanistic and Technical-Professional)
- Type of high school (Private, Municipal, Subsidized)
- Type of university (High, Medium, and Low admission scores)

RESULTS

Phase 1 Results

This section presents the results of the alignment study and is organized by test with the alignment results for each of the four Webb analyses for the curriculum framework as well as standards from the Scientific-Humanistic and Technical-Professional curricular branches.

PSU Language and Communication Exam Results

Categorical Concurrence
For the purposes of this alignment, at least six test items must align per content standard for sufficient categorical concurrence.

See Table 119 for summary results by standard set. The analysis were carried out considering the differentiated formation (Scientific-Humanistic and Technical-Professional) and the common contents from first through fourth grade high school for both modes. The table shows two standard sets. One pertains to Differentiated Training and the other to Curriculum Framework (i.e., *Formación General*). The Differentiated Training depicts the branch of Chile's national curriculum, which is differentiated into Scientific-Humanistic and Technical-Professional areas. The Curriculum Framework depicts the scope for the whole national high school curriculum that is expected for all high school students in Chile. As shown in Table 119, the minimal target of at least six items aligning per content standard was not met for either standard set (*Differentiated Training* (DT) Scientific-Humanistic*)* or Curriculum Framework (CF). For the DT standard set, one of the standards was not addressed by a single assessment item. This was also the case for eight of the 16 standards in the CF standard set. (Note: for CF the alignment study rolled up the OF as one standard and the three CMO – oral communication, reading and writing, as three separate standards for a total of four standards. Since CF spans over the four grades of high school, the number of standards become sixteen. For DT the alignment study relied on four standards: Language and Society (OF), Language and Society (CMO), Literature and Identity (OF) and Literature and Identify (CM).

Table 119: Categorical Concurrence of the Language and Communication Exam

Standard set	# of Standards	Mean # of items per standard	# of standards aligned to 6 items	Findings
Differentiated Training (SH)	4	3	1	Low
General Curriculum Framework (SH and TP)	16	31.5	8	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Depth-of-knowledge Consistency

In order to be deemed acceptable based on depth-of-knowledge criteria, at least 50% of the PSU assessment items must match or exceed the level of intellectual complexity required by the objective which they address. Table 120 shows that the Language and Communication exam does not reach this benchmark, with 10% of assessment items in being as cognitively demanding as the objective in either standard set.

Table 120: DOK Consistency of the Language and Communication Exam

Standard set	Mean DOK score	# of items at or above mean DOK	# of items below mean DOK	Findings
Differentiated Training (SH)	2.7	8	72	Low
General Curriculum Framework (SH and TP)	2.8	8	72	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Range-of-knowledge Correspondence

As described above, range-of-knowledge correspondence measures alignment of the breadth of knowledge between the content strands and the assessment. In order to be considered sufficient, 50% or more of the objectives within a standard must align to one or more assessment item. See Table 121 for summary results. The Language and Communication exam has a range-of-knowledge correspondence of below 40% for both DT and CF standard sets, and thus is deemed to be unacceptable for this criterion.

Table 121: Range-of-knowledge Correspondence of the Language and Communication Exam

Standard set	# of standards	# standards in which 50% of objectives align with an assessment item	Findings
Differentiated Training (SH)	4	0 (0%)	Low
General Curriculum Framework (SH and TP)	16	5 (31.3%)	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Balance of Representation

Balance of representation measures the distribution of tasks across the objectives within a standard with the goal of an even distribution. Balance of representation is calculated according to Webb's balance index (Webb, 2005). The cutoff for a sufficient level of balance is a balance index score of at least 0.7. The results for the balance of representation analysis are presented in Table 122. As the PSU Language and Communication exam scored below an index value of 0.7 for both DT and CF standard sets, it is deemed unacceptable.

Table 122: Balance of Representation of the Language and Communication Exam

Standard set	Balance index	Findings
Differentiated Training (SH)	0.25	Low
General Curriculum Framework (SH and TP)	0.28	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Categorical Concurrence

For the purposes of this alignment, at least six test items must align per content standard for sufficient categorical concurrence. As shown in Table 123, this minimal target was not met for either standard set (Differentiated Training (DT) Scientific-Humanistic – Mathematics or Curriculum Framework (CF) – Mathematics sector) by the Mathematics exam. For the DT standard set, five of the standards were not addressed by a single assessment item. This was also the case for 9 of the 19 standards in the CF standard set.

Table 123: Categorical Concurrence of the Mathematics Exam

Standard set	# of Standards	Mean # of items per standard	# of standards aligned to 6 items	Findings
Differentiated Training (SH)	8	1.75	1	Low
General Curriculum Framework (SH and TP)	19	2.26	2	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Depth-of-knowledge Consistency

In order to be deemed acceptable based on depth-of-knowledge criteria, at least 50% of the PSU assessment items must match or exceed the level of intellectual complexity required by the objective which they address. Table 124 shows that the Mathematics exam falls below this benchmark.

Table 124: DOK Consistency of the Mathematics Exam

Standard set	Mean DOK score	# of items at or above mean DOK	# of items below mean DOK	Findings
Differentiated Training (SH)	2.5	24	51	Low
General Curriculum Framework (SH and TP)	2.4	24	51	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Range-of-knowledge Correspondence

As described above, range-of-knowledge correspondence measures alignment of the breadth of knowledge between the content strands and the assessment. In order to be considered acceptable, 50% or more of the objectives within a standard must align to one or more assessment item. The Mathematics exam has a range-of-knowledge correspondence of below 40% for both DT and CF standard sets, and thus is deemed to be below this criterion (Table 125).

Table 125: Range-of-knowledge Correspondence of the Mathematics Exam

Standard set	# of standards	# standards in which 50% of objectives align with an assessment item	Findings
Differentiated Training (SH)	8	2 (25%)	Low
General Curriculum Framework (SH and TP)	19	4 (21.1%)	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Balance of Representation

Balance of representation measures the distribution of tasks across the objectives within a standard with the goal of an even distribution. Balance of representation is calculated according to Webb's balance index (Webb, 2005). The cutoff for an "acceptable" level of balance is a balance index score of at least 0.7. The results for the balance of representation analysis are presented in Table 126. The PSU Mathematics exam scored below an index value of 0.7 for both DT and CF standard sets.

Table 126: Balance of Representation of the Mathematics Exam

Standard set	Balance index	Findings
Differentiated Training (SH)	.19	Low
General Curriculum Framework (SH and TP)	.20	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Categorical Concurrence

For the purposes of this alignment, at least six test items must align per content standard for sufficient categorical concurrence. As shown in Table 127, this minimal target was not met for either standard set (Differentiated Training (DT) Scientific-Humanistic – History and Social Sciences or Curriculum Framework (CF) – History, Geography and Social Sciences sector) by the History and Social Science exam. For the DT standard set, two of the standards were not addressed by a single assessment item. All of the CF standards were addressed by at least one assessment item.

Table 127: Categorical Concurrence of the History and Social Science Exam

Standard set	# of Standards	Mean # of items per standard	# of standards aligned to 6 items	Findings
Differentiated Training (SH)	8	12	3	Low
General Curriculum Framework (SH and TP)	8	10.9	6	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Depth-of-knowledge Consistency

In order to be deemed acceptable based on depth-of-knowledge criteria, at least 50% of the PSU assessment items may match or exceed the level of intellectual complexity required by the objective which they address. Table 128 shows that the History and Social Science exam does not reach this benchmark, with 43% of assessment items being as cognitively demanding as the DT standards and 17% of the assessment items being as cognitively demanding as the CF standards.

Table 128: DOK Consistency of the History and Social Science Exam

Standard set	Mean DOK score	# of items at or above mean DOK	# of items below mean DOK	Findings
Differentiated Training (SH)	1.4	32	43	Low
General Curriculum Framework (SH and TP)	2.6	13	62	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Range-of-knowledge Correspondence

As described above, range-of-knowledge correspondence measures alignment of the breadth of knowledge between the content strands and the assessment. In order to be considered "acceptable," 50% or more of the objectives within a standard may align to one or more assessment item. The History and Social Science exam has an acceptable range-of-knowledge correspondence for both standard sets (Table 129).

Table 129: Range-of-knowledge Correspondence of the History and Social Science Exam

Standard set	# of standards	# standards in which 50% of objectives align with an assessment item	Findings
Differentiated Training (Scientific-Humanistic)	8	5 (62.5%)	High
Curriculum Framework (SH and TP)	8	5 (62.5%)	High

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

Balance of Representation

Balance of representation measures the distribution of tasks across the objectives within a standard with the goal of an even distribution. Balance of representation is calculated according to Webb's balance index (Webb, 2005). The cutoff for an "acceptable" level of balance is a balance index score of at least 0.7. The results for the balance of representation analysis are presented in Table 130. As the PSU History and Social Science exam scored below an index value of 0.7 for both DT and CF standard sets, it is deemed unacceptable.

Table 130: Balance of Representation of the History and Social Science Exam

Standard set	Balance index	Findings
Differentiated Training (SH)	0.36	Low
General Curriculum Framework (SH and TP)	0.47	Low

(Note: SH = Scientific-Humanistic; TP = Technical-Professional)

PSU Science Exams Results

The results for the Webb alignment of the individual Science exams (Biology, Chemistry, and Physics) are presented below. These exams were aligned individually to the corresponding DT standards. The Natural Sciences Curriculum Framework standard set was divided into subject-based subsets (Biology, Chemistry, and Physics), and then the exams were aligned to their corresponding sub-sets. The Chemistry exam was also aligned to the Technical-Professional (TP) Differentiated Training - Chemistry sector standard set. In addition to doing these specialized Science standard alignments, we also pooled the assessment items from the three Science tests and aligned them to the complete Natural Sciences Curriculum Framework standard set. The results from this alignment are presented following those from the more focused alignments.

Categorical Concurrence – Individual Science exams
For the purposes of this alignment, at least six test items must align per content standard for sufficient categorical concurrence. As shown in Table 131, this minimal target was not met for any of the three Science exams with any of the standard sets.

Table 131: Categorical Concurrence of the Biology, Chemistry, and Physics Exams

Test	Standard set	# of Standards	Mean # of items per standard	# of standards aligned to 6 items	Findings
Biology	DT Biology	6	2	0	Low
	CF Natural Sciences – Biology	16	6.3	8	Low
Chemistry	DT Chemistry	6	0	0	Low
	CF Natural Sciences – Chemistry	12	4.9	6	Low
	TP Chemistry	4	0	0	Low
Physics	DT Physics	9	2.6	2	Low
	CF Natural Sciences – Physics	18	2.8	4	Low

Depth-of-knowledge Consistency – Individual Science exams
In order to be deemed acceptable based on depth-of-knowledge criteria, at least 50% of the PSU assessment items may match or exceed the level of intellectual complexity required by the objective which they address. Table 132 shows the Chemistry exams falls below that threshold when aligned with both the CF Natural Sciences – Chemistry standards and the TP Chemistry standards. When aligned to the DT Physics standards, the Physics exam is also rated as "below the cut."

Table 132: DOK Consistency of the Biology, Chemistry, and Physics Exams

Test	Standard set	Mean DOK score	# of items at or above mean DOK	# of items below mean DOK	Findings
Biology	DT Biology	2.0	31	13	High
	CF Natural Sciences – Biology	1.8	31	13	High
Chemistry	DT Chemistry	1.5	31	13	High
	CF Natural Sciences – Chemistry	2.2	1	43	Low
	TP Chemistry	2.3	1	43	Low
Physics	DT Physics	2.3	5	39	Low
	CF Natural Sciences – Physics	2.0	39	5	High

Range-of-knowledge Correspondence – Individual Science exams
As described above, range-of-knowledge correspondence measures alignment of the breadth of knowledge between the content strands and the assessment. In order to be considered "acceptable," 50% or more of the objectives within a standard should align to one or more assessment item. None of the three Science exams reached that expectation, regardless of the standard set to which they were aligned (Table 133).

Table 133: Range-of-knowledge Correspondence of the Biology, Chemistry, and Physics Exams

Test	Standard set	# of standards	# standards in which 50% of objectives align with an assessment item	Findings
Biology	DT Biology	6	3	Low
	CF Natural Sciences – Biology	16	3	Low
Chemistry	DT Chemistry	6	0	Low
	CF Natural Sciences – Chemistry	12	1	Low
	TP Chemistry	4	0	Low
Physics	DT Physics	9	2	Low
	CF Natural Sciences – Physics	18	1	Low

Balance of Representation – Individual Science Exams
Balance of representation measures the distribution of items across the objectives within a standard with the goal of an even distribution. Balance of representation is calculated according to Webb's balance index (Webb, 2005). The cutoff for an "acceptable" level of balance is a balance index score of at least 0.7. The results for the balance of representation analysis are presented in Table 134. All three of the Science exams had balance indexes below 0.7, and hence their balance representation is deemed low.

Table 134: Balance of Representation of the Three Science Exams (Biology, Chemistry, and Physics).

Test	Standard set	Balance index	Findings
Biology	DT Biology	0.30	Low
	CF Natural Sciences – Biology	0.30	Low
Chemistry	DT Chemistry	N/A (no alignments)	Low
	CF Natural Sciences – Chemistry	0.30	Low
	TP Chemistry	N/A (no alignments)	Low
Physics	DT Physics	0.26	Low
	CF Natural Sciences – Physics	0.25	Low

Categorical Concurrence – Pooled Science Exams
For the purposes of this alignment, at least six test items must align per content standard for sufficient categorical concurrence. As shown in Table 135, this minimal target was not met when the pooled Science exams were aligned with the entire Curriculum Framework (CF).

Table 135: Categorical Concurrence of the Pooled Science Exams

Standard set	# of Standards	Mean # of items per standard	# of standards aligned to 6 items	Findings
CF –Natural Sciences	46	4.6	17	Low

Depth-of-knowledge Consistency – Pooled Science exams

In order to be deemed acceptable based on depth-of-knowledge criteria, at least 50% of the PSU assessment items must match or exceed the level of intellectual complexity required by the objective which they address. Table 136 shows that the pooled exam does meet this threshold.

Table 136: DOK Consistency of the Pooled Science Exams

Standard set	Mean DOK score	# of items at or above mean DOK	# of items below mean DOK	Findings
CF – Natural Sciences	2.0	101	31	Low

Range-of-knowledge correspondence – Pooled Science exams

As described above, range-of-knowledge correspondence measures alignment of the breadth of knowledge between the content strands and the assessment. In order to be considered "acceptable," 50% or more of the objectives within a standard must align to one or more assessment item. The pooled Science exam did not meet this level when aligned with the CF – natural Sciences standard set (Table 137).

Table 137: Range-of-knowledge Correspondence of Pooled Science Exams

Standard set	# of standards	# standards in which 50% of objectives align with an assessment item	Findings
CF – Natural Sciences	46	5 (10.9%)	Low

Balance of Representation – Pooled Science exams

Balance of representation measures the distribution of items across the objectives within a standard with the goal of an even distribution. Balance of representation is calculated according to Webb's balance index (Webb, 2005). The cutoff for an "acceptable" level of balance is a balance index score of at least 0.7. The pooled Science exam did not reach this expectation.

Table 138: Balance of Representation of the Pooled Science Exams

Standard set	Balance index	Findings
CF –Natural Sciences	0.29	Low

Phase 2 Results

For Phase 2, panels were assembled after the relevant participants were identified by Chile's Ministry of Education and recruited by Pearson. When feasible, one day before each panel meeting, Pearson electronically distributed the goals of the meeting and the discussion protocols that were to be used so that panel members could prepare. Panels of professionals met separately.

The interviewers met with participants that represented two different groups of PSU stakeholders: university teachers and high school teachers.

Eleven university teachers were interviewed. They were all from state, metropolitan universities that were part of CRUCH. There was one teaching director, four Mathematics professors, one Physics professor, two Chemistry professors, two Biology professors, and one professor of Language and Communication.

The high school teachers represented the largest group of participants that were interviewed. The 16 teachers came from two private, metropolitan high schools that followed a Scientific-Humanistic curricular branch. Among the teachers were four that specialized in Language and Communication, four in Mathematics, three in History and Social Sciences and five in Science. The Science teachers consisted of two Chemistry teachers, two Physics teachers and one Biology teacher.

The protocols that were used to interview the participants and guide their discussion varied by group and are presented here:

Pauta Entrevista: **Profesores de universidad**

Instrucciones: Familiarícese con los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) del marco curricular de la enseñanza media para esta asignatura, así como el marco curricular para la formación diferenciada, los cuales se encuentran ubicados en su cuaderno de trabajo. Mientras lea la documentación, por favor considere y responda a las siguientes preguntas:

1. Utilizando los OF y los CMO, haga una marca de verificación o checkmark (√) para identificar esos objetivos que en su opinión más fuertemente definen las características del conocimiento que deben poseer al principio de su enseñanza universitaria.

2. Dibuje un círculo alrededor de esos OF y CMO en los cuales usted se sienta que los estudiantes tienen las mayor dificultades. Ofrezca sus pensamientos de por qué esto podría ser el caso.

3. ¿Hasta qué punto siente usted que la PSU como instrumento de selección universitaria, indica el nivel de preparación de los estudiantes universitarios al comienzo de sus estudios?

4. Escriba un breve resumen del consenso grupal.

5. Por favor proporcione otros comentarios y/o experiencias que usted desee compartir.

Pauta Entrevista: **Profesores de secundaria**

Instrucciones: Familiarícese con los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) del marco curricular de la enseñanza media para esta asignatura, así como el marco curricular para la formación diferenciada, los cuales se encuentran ubicados en su cuaderno de trabajo. Mientras lea la documentación, por favor considere y responda a las siguientes preguntas:

1. Utilizando los OF y los CMO, haga una marca de verificación o checkmark (√) para identificar esos objetivos que en su opinión están cubiertos en las aulas de su área temática.

2. Dibuje un círculo alrededor de esos OF y CMO en los cuales usted se sienta que los estudiantes tienen las mayor dificultades. Ofrezca sus pensamientos de por qué esto podría ser el caso.

3. Escriba un resumen en alto nivel del consenso grupal.

4. Por favor proporcione otros comentarios y/o experiencias que usted desee compartir.

SUMMARY OF INTERVIEWS

The interview findings and discussions have been summarized below in Table 139 through Table 141 of the evaluation report. Each table summarizes the results from all three of the stakeholder groups (university teachers and high school teachers) for a single topic.

Table 139 provides a summary of the interviewees' comments regarding the degree of alignment of the PSU test domain to that of the high school content. A major theme that came from the stakeholders was that there was a "dissociation between the PSU and the school academic content," with the alignment being especially bad for Science. The alignment was also seen to be worse for those students taking the Technical-Professional curricular branch than for those taking the Scientific-Humanistic curricular branch.

Table 139: Summary of Discussion Protocols for Degree of Alignment

Degree of alignment of PSU domain to high school content	
Major Points	**Minor Points**
University Teachers	
• There is no symmetry in terms of what is being implemented in terms of basic and fundamental content. • There is a disconnect between what the PSU measures and academic objectives.	• PSU incorrectly regarded as definition of quality of a student.
High School Teachers	
• Dissociation between school content and PSU. PSU does not assess fundamental objectives of demonstrating, conjecturing, and explaining. • With respect to Science, the PSU places too much emphasis on Science content and not enough on scientific reasoning skills. • The training at Scientific-Humanistic schools more closely aligns with the content of the PSU and with what is taught in colleges than the training at professional/technical schools.	.

Table 140 presents the interviewees' comments and discussion regarding the use of the PSU as a prediction of success in higher education. A major theme here was that the PSU does not capture all of the knowledge and skills needed to do well in higher education. The curriculum experts, for example, stated while they considered the alignment of the PSU to the CMO to be high, the PSU tested students at a lower level of ability (such as recall) when a higher level of ability may be more appropriate for university success. Both the curriculum specialists and the university teacher stated that PSU tests do not capture student

motivation or other important qualities and that students who do poorly on the PSU tests may go on to be successful at the university. Other themes that were captured included the perception that students with lower socio-economic status (SES) were at a disadvantage on the PSU tests and that teachers in 11th and 12th grades focus more on teaching to the PSU tests than to teaching the curriculum.

Table 140: Summary of Discussion Protocols for PSU Prediction of Success in Higher Education

The battery of tests that make up the PSU bases its construction in the CMO Curriculum Framework of School. The current selection system means that students who receive the highest scores represent the best chance of success in fulfilling the tasks required in higher education. Do you agree with this statement or not? Please explain your answer	
Major Points	**Minor Points**
University Teachers	
• Scores seem to decrease as moved away from capitol. Suggested this was political/social problem as schools further away from the capitol tended to be lower SES. • The perception of test is that it is unfair – low SES students are at a disadvantage. • Indicated the PSU did not capture motivation. Gave example of a student that performed poorly on a test, but through motivation, achieved success. • Too much emphasis for a single test. • Test does not capture all qualities that contribute to college success. • PSU is deficient in that it does not measure knowledge and skills need for success in Higher Education.	• Suggested adding class rank as a selection criterion. Pointed to its successful implementation in Texas and Peru. Noted it was not conflated with SES and school type the way PSU scores were. (Note: Independent research supports the notion (Baron & Frank; 1992; Niu & Tienda, 2009). • Suggested administration of PSU should be done at the macro/state level and not by the universities. • Did not think problems with PSU were technical. Thought it was well constructed, but flawed in other ways.
High School Teachers	
• PSU does not cover all relevant content. Aspects important to success in college that is not covered by PSU. • Noted the importance of Science and that there were only 18 Science items on the PSU. • Observed that the PSU does not assess the fundamental objective of demonstrating, conjecturing, explaining. Neither does it assess reliability tables.	• Language and society portion employs a linguistic approach (i.e. different types of clear language) that is not of value to students, and is not of practical use in college. • Pointed out that the nature of the PSU – a multiple choice test – limited the extent to which it could assess higher order scientific skills.

The summary in Table 141 describes the interviewees' responses and discussions on the relationship between the high school curriculum and the level of knowledge needed to be successful at University. Here again, the earlier theme of the PSU testing students, at a low level of skill rather than at a high level of skill, is repeated. Specific areas of deficiency are seen in Science and the lack of statistics training in Mathematics. The point about teaching to the PSU test comes up again.

Table 141: Summary of Discussion Protocols for Relationship between High School Curriculum and Level of Knowledge

On relationship between Chile's national high school curriculum and level of knowledge required for entry level students to be successful	
Major Points	**Minor Points**
University Teachers	
• There is no symmetry in terms of what is being implemented in terms of basic and fundamental content. • Curriculum focuses on recall rather than comprehension and application, particularly in science and math.	
High School Teachers	
• Noted that schools employing a Scientific-Humanistic approach more adequately prepared students for college than programs with a Technical-Professional approach. • There are aspects of the CMO that are covered superficially. Instead, emphasis is placed on preparing students for the PSU.	• High school students typically receive insufficient training in demonstrations and conjectures. • Most of the CMOs in Science are addressed in high school, but not in the depth and meaningfulness desired. • True with language arts/reading comprehension/oral communication. Essentially address all CMOs and OFs in high school, but later in 12 grade, veer away to prepare students for the PSU.

SUMMARY OF STATEMENT FROM THE CURRICULUM UNIT

The Curriculum Unit of MINEDUC has offered two documents to the evaluation team that provide perspective on the National Curriculum of Chile and its relation to the PSU. Those documents are provided as appendices to this evaluation report.

In response to a July 2009 request from DEMRE, the Curriculum Unit analyzed the PSU assessment frameworks in relation to the curriculum. See the document, *Revisión de Marcos Teóricos de Evaluación para PSU*, included in Appendix Y of this report. This review concluded that the PSU does not achieve an appropriate curricular reference because it is concentrated upon the Minimum Obligatory Contents (CMO) and not upon the Fundamental Objectives (OF), which are the nucleus of the National Curriculum. One of the recommendations of this report was to strengthen the relation between the Curriculum Unit and DEMRE. The joint work would allow for a more faithful interpretation of the National Curriculum and an improved prioritization of the educational objectives of the National Curriculum that are targeted by the PSU.

The Curriculum Unit met with DEMRE in 2010 to establish working groups to address the propositions found in the aforementioned 2009 report. The challenges of the PSU addressing the ongoing curricular changes were addressed by the Curriculum Unit in these meetings. A summary of the key themes of these meetings is included in a PowerPoint presentation, *Ajuste Curricular – PSU*, incorporated into Appendix Z of this report. There was no follow-up to these meetings by DEMRE, so it is unknown if they took action. What is now contained on the website of this institution shows that there has been no progress in this regard.

The following are the principal concerns of the Curriculum Unit with respect to the PSU. These have been communicated to DEMRE as well as to the evaluation team.

- Consistent Alignment: the main concern, which derives from the report requested by DEMRE itself, is that the PSU evaluation framework uses as reference one section of the curriculum, the CMO, which do not necessarily render account of the total extent of the curriculum. In other words, the PSU utilizes a methodology which aligns it only superficially with the curriculum, leaving aside that which is central and the fundamental aspects of the curriculum.

- The Chilean Curriculum, as any modern curriculum, has recently been updated and revised. Specifically, the Secondary Educational curriculum was modified significantly in 2009: This change has been implemented gradually year by year. Therefore the existence of certainty and transparency is fundamental, for the whole educational system, in relation to which curriculum is being evaluated and how the intersections between two curricula are constructed.

- To date, 45% of secondary school enrollment corresponds to the Technical-Professional curricular branch. An increasing number of graduates of this curricular branch takes the PSU as part of the admissions process into higher education. From Curriculum Unit's perspective, there is a concern about the distance between that which is declared (the PSU as a general assessment) and that which is real (the PSU as a general and differentiated assessment, which emphasizes the Scientific-Humanistic curricular branch). [MINEDUC, personal communication, January 2013]

EVALUATION

This study made use of two kinds of analyses to examine content validity of the PSU tests.

The Phase 1 analysis made use of the alignment methodology of Webb (1997). Four alignment criteria were used to examine the PSU tests: 1) categorical concurrence, which addresses whether the assessment covers the same broad content categories, 2) depth-of-knowledge consistency, which compares the level of cognitive complexity required by the standards with those required by the test items, 3) range-of-knowledge correspondence, which examines the breadth of knowledge assessed compared to the expectations set forth in the standards, and 4) balance of representation, which examines the distribution of assessment items across content objectives.

The results of the Webb analysis indicate that across all of the criteria listed for almost all of the PSU tests, the level of alignment of the PSU to both the Fundamental Objectives (OF) and Mandated Minimal Contents (CMO) of the Chilean curriculum was uniformly low. One aspect to take into account when interpreting these results is that there are certain strands within the standard sets that are impossible to assess in a multiple-choice format exam of the PSU. This would tend to lower the scores for all Webb criteria except for depth-of-knowledge correspondence. Examples of such strands include "oral communication" and "writing" from the Curriculum Framework – Language and Communication sector and the majority of the objectives from the Technical-Professional curricular branch – Chemistry sector as they address skills that could only be assessed in a laboratory setting. While the existence of these strands within the standards will result in artificially low scores for the PSU tests using the Webb method, the results demonstrate that much improvement could be made.

The Phase 2 analysis made use of interviews and discussions with three important PSU stakeholder groups, namely, curriculum specialists, university teachers and high school teachers. The interviews and discussions focused on three broad issues: 1) The degree alignment of the PSU domain to high school classroom instruction; 2) whether the PSU as a university selection test adequately indicates readiness of entry-level university students at the beginning of university instruction; 3) the relationship between Chile's national high school curriculum and level of knowledge required for entry-level university students to be successful.

The results of the interviews confirm and extend the Webb analysis results. In the interviewees' opinion, the PSU tests tended to focus on rote knowledge rather than on the application of knowledge or qualities such as motivation that would be needed for students to be successful at university. They were also concerned with the effects of teaching to the PSU tests rather than to the national curriculum and what they felt were the disadvantages faced by low SES or Technical-Professional curricular branch students when taking the PSU tests.

RECOMMENDATIONS

1. There is a fundamental disconnect between the purpose and use of the PSU to select students for university admissions and the content of the PSU that is based on Chile's high school Curricular Framework (*Marco Curricular*). We recommend a review of the policy of using the Curricular Framework as the basis for the development of the PSU test frameworks. As a part of this review, we recommend the development of a framework that describes the aptitudes (i.e., abilities) and relevant non-cognitive variables (e.g., study skills and motivation) needed by students in order to be successful at the university. Such a framework would focus the PSU on the aptitudes necessary to succeed at the university and complement the measure of high school achievement found in NEM and combined together in the postulation score.
2. Although we have recommended aligning the PSU tests to standards for success at the university, we acknowledge the urgency to develop the 2013-14 test forms a based on the full implementation of the 2009 curricular reform. To that end, we recommend performing an alignment study on these PSU test forms. The results of this study should inform the broader recommendation to redirect the emphasis of the PSU to university success.
3. We recommend reviewing the item types used on the PSU tests to address the perceived low level of cognitive complexity found on the tests due to the exclusive use of multiple-choice items.

BIBLIOGRAPHY

Baron, J., & Frank, N. (1992). SATs, achievement tests, and high school class rank as predictors of college performance. *Educational and Psychological Measurement, 52*, 1047-1056.

Hess, K. K. (2005). *Cognitive complexity: Applying Webb DOK levels to Bloom's taxonomy*. Dover, NH: National Center for Assessment.

Kobrin, J. L., Patterson, B. E., Shaw, E. J. Mattern, K. D., & Barbutt, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average.* College Board Research Report No. 2008-5.

Niu, S. X., & Tienda, M. (2009). *Testing, ranking and college performance: Does high school matter?* Unpublished manuscript.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* [Research Monograph No. 6]. Washington, D.C.: Council of Chief State School Officers.

Webb, N. L. (2005). *Webb alignment tool: Training manual*. Madison, WI: Wisconsin Center for Education Research. Available: http://www.wcer.wisc.edu/WAT/index.aspx.

Objective 2.3. Analysis of trajectories of PSU scores for subpopulations throughout time, considering dependence, mode and gender

ABSTRACT

The primary focus of this research was to analyze trajectories of PSU scores from 2004 to 2011 and identify variables that moderated those trajectories. Trends from 2004 to 2006 and from 2007 to 2011 were examined in order to determine if modifications to the test affected the scores. A second objective of this research was to identify school-level characteristics that moderated the degree of covariation between PSU subtest scores and a measure of high school performance. The study relied on longitudinal data sets spanning from 2004-2011. DEMRE provided databases for PSU test scores and high school GPA (NEM), and MINEDUC provided databases with university outcomes.

Results of the trend analysis indicate that, on average, PSU scores remained fairly consistent over time with a slight upward trend beginning in 2007. An examination of subpopulations indicates that this upward trend is largely due to the performance of private schools and schools with a Scientific-Humanistic curricular branch. Trend lines disaggregated by school type and curricular branch showed that scores steadily increased over time for private schools and schools with a Scientific-Humanistic curricular branch, while the scores stayed flat for public and technical schools. In addition to differences in scores due to school type and curricular branch, gender, socioeconomic status (SES) and the region in which the student resided significantly moderated the trend of PSU scores. Not unexpectedly, SES was strongly related to PSU scores. This is in-line with previous research on the subject that demonstrates a moderate to strong relationship between SES and academic achievement (Sirin, 2005; White, 1982). Males tended to perform better on PSU tests in most subjects despite typically having lower NEM. Students in the central region typically scored higher on the PSU than their counterparts in the northern and southern regions of the country.

A secondary focus of this research was to examine the covariation between high school performance grade point average and PSU scores. It was discovered that, while NEM predicted performance on all PSU subtests, it performed particularly well for Mathematics and Science. School-level variables that moderate this relationship between NEM and PSU test scores were also examined. School type and curricular branch were particularly strong moderators of this relationship as the slope for NEM was substantially steeper for private schools and schools providing the Scientific-Humanistic curricular branch. SES, region, and the percentage of females at a high school also moderated the relationship between NEM and PSU scores.

INTRODUCTION

Internationally, admission decisions for examinees to colleges and universities involve the use of admissions test data. An analysis of the trajectory of university admissions test scores through time has become an important element of institutional validity studies because such longitudinal analyses aid in the identification of gaps in test performance over time for relevant subpopulations. When performing such analyses, test scores are reported on scales that are maintained through equating processes to control the changing difficulty of admission tests across years. In this respect, differences in scores across time that are related to the variability of cohort performance are comparable.

In Chile, the PSU test battery was introduced as part of the 2004 admission process, and it has been in place through the 2012 admission process. Over this time, the PSU has undergone a number of changes and has had a number of constraints associated with it.

- The PSU has undergone several changes in its assessment frameworks as a result of policy recommendations and curriculum reforms. From 2004 to 2006 there was a gradual inclusion in the PSU of certain content specified in the national high school curriculum.
- Second, the PSU's intended population of test takers has changed dramatically as a result of an increase of applicants graduating from high schools with Technical-Professional orientations. This increase started during the 2006 admission process as a result of modifications to Chile's *Ley Orgánica Constitucional de la Enseñanza,* when high school graduates from municipal schools were granted free PSU registration (LOCE, Articulo 8).
- Third, the PSU scale has not been maintained across time using equating methodology. As a result, year-to-year differences in form difficulties have not been accounted for, and they are confounded with cohort differences.
- Fourth, beginning 2005 admission process, the PSU reported a single score for Science. For previous admission processes Science scores were reported separately for Biology, Physics, and Chemistry.
- Fifth, PSU test length has increased as a result of intended and unintended changes. Among intended changes is the increase in length for the PSU Mathematics test for 2012 admission process. The test length was increased by adding five items tapping on upper regions of the difficulty scale. Among unintended changes is the voiding of items that happen as a result of conflicting item keys found after operational administration of the tests.

METHODOLOGY

When analyzing change in test scores for major subpopulations over a period of time, different techniques are available for use, depending on the purpose of the analysis. For criterion referenced testing (e.g., achievement testing), analyses center on describing achievement performance levels variability by disaggregated subpopulations. For norm referenced testing (e.g., university admissions testing), the analysis often involves examining average scores for subpopulations of interest over time.

The evaluation team conducted trend analyses on PSU standardized scores utilizing student data sets from 2004 to 2011. Data for all students who took the PSU were used, regardless of whether they enrolled in a post-secondary institution. In addition, the evaluation team analyzed two subgroups of data sets: 2004 to 2006 and 2007 to 2011.[11]

The evaluation team researched the following questions:

- What is the trend in mean PSU scores by PSU subtest for the following subpopulations?
    - Gender: Male or Female
    - Region: North (codes 1, 2, 3, 4, 15), Central (5, 13 [RM]), or South (6, 7, 8, 9, 10, 11, 12, 14)
    - Socio-economic status[12]: Low (quintile A), Below Mean (quintile B), Average (quintile C), Above Average (quintile D), High (quintile E)
    - Curricular Branch: Scientific-Humanistic or Technical-Professional
    - Type of high school: Private, Subsidized, or Municipal
- What school-level variables moderate the relationship between PSU scores and NEM?

Descriptive and Inferential Statistical Analysis

In order to address the first research question, descriptive statistics were generated for PSU scale scores by each admission year for the total population and for the subpopulations mentioned above. Descriptive information such as n-counts, mean, and standard deviation were computed, reported and interpreted. As part of the descriptive statistics, plots of mean scale scores (with 95% confidence bands around mean scale scores) were generated and interpreted.

Hypothesis testing involving an analyses of variance factorial design with one dependent variable was carried out to test for null differences on PSU mean scale scores by year and by subpopulation. That is, the ANOVA design for each dependent variable looked at a main effect for year, a main effect for subpopulation and an interaction term for year-by-subpopulation.

For the second research question, the effects of school-level characteristics on levels of covariation between PSU subtest scores and NEM were identified. To accomplish this, the evaluation team employed hierarchical linear modeling (HLM) (Kreft & De Leeuw, 1998; Raundenbush & Bryk, 2002). The model allowed us to investigate extent to which school-level characteristics (i.e. type, curricular branch) affected the relationship between NEM and

[11] The evaluation team also inspected the trends of PSU raw scores. Selected results are show in this objective; additional results are documented in Appendix L, Appendix M, and Appendix N.

[12] The process for computing SES quintiles used a combination of family income and mother's educational attainment.

PSU scores. The variables examined were limited to the specific research question and included:

- Criterion: PSU scores
- Level 1 unit of analysis: student
- Student-level predictor: NEM
- Level 2 unit of analysis: high school
- School-level predictors: school-level SES, curricular branch, school type, region, and % of student population that is female.

As 0 was not a valid value for NEM, grand mean centering was implemented in order to facilitate interpretation of results. School-level SES was calculated by averaging the student-level SES within each school in a process similar to the one illustrated in Bryk & Raudenbush (1992). School-level SES was centered as well for the same reason NEM was; namely, 0 was not within the range of values for SES. The percentage of females ranges from 0 to 1 and estimates the percentage of female students within a school.

Also, the relationship between high school performance (NEM) and PSU scores was examined using Pearson product moment correlations within each high school (r_i). These correlations were then aggregated into a sample weighted correlation using the procedure outlined by Hunter and Schmidt (1990), which relies on school sample size (N_i).

$$\text{Weighted } r = \sum(N_i * r_i)/\sum N_i$$

RESULTS
Part I

Part I involved identifying trends in PSU scores and variables that moderated those trends. The first step was to generate descriptive statistics for all subtests of the PSU. PSU scale score and NEM results are presented in Table 142 and raw score statistics are presented in Table 143. The descriptive statistics for all subtests appeared reasonable.

There are important patterns present in the following table that require more discussion. First, there is a general increase in the number of applicants across year which is consistent with the history of the implementation of the testing program. Second, the mean and standard deviation of the scale scores systematically depart from the PSU scale (mean=500 and SD=110). This result is consistent with the practice of truncation of the PSU scale as implemented by DEMRE. Finally, the kurtosis and skewness coefficients indicate that the distributions depart from normality for most years and subtests.

Note that the separate scale score transformation for NEM results (mean=500 and SD=100) in a scale that has substantially higher means and distinctly non-normal distributions. The standard deviations for the NEM results are lower than the PSU standard deviations because of the difference in scaling constants used to construct the distributions (i.e., SD=100 versus SD=110).

Table 142: Descriptive Statistics for PSU Scale Scores and NEM by Year and Subtest

Year	Subject	N	Mean	S.D.	Kurtosis	Skew	Min	Max
2004	Language and Communication	128705	491.14	122.30	-0.73	0.25	175	840
2005	Language and Communication	128284	490.13	112.43	-0.22	0.09	150	850
2006	Language and Communication	132944	489.98	112.73	-0.24	0.10	150	850
2007	Language and Communication	163049	490.26	111.96	-0.20	0.11	150	850
2008	Language and Communication	164927	492.90	112.44	-0.20	0.08	150	850
2009	Language and Communication	185930	496.08	113.20	-0.26	0.07	166	850
2010	Language and Communication	200520	497.42	112.49	-0.27	0.05	165	850
2011	Language and Communication	187819	493.99	111.21	-0.29	0.12	150	850
2004	Mathematics	128705	490.47	110.50	0.30	0.15	112	840
2005	Mathematics	128284	489.94	113.11	-0.02	0.19	150	850
2006	Mathematics	132944	490.97	113.27	-0.01	0.21	150	850
2007	Mathematics	163049	490.81	112.07	-0.07	0.16	150	850
2008	Mathematics	164927	493.61	113.27	-0.08	0.15	150	850
2009	Mathematics	185930	496.47	114.29	-0.17	0.10	150	850
2010	Mathematics	200520	499.18	114.94	-0.03	0.14	157	850
2011	Mathematics	187819	495.36	113.98	0.12	0.22	150	850
2004	Science	67491	493.30	101.50	-0.02	0.07	130	835
2005	Science	68134	490.16	112.40	-0.14	0.11	150	850
2006	Science	72097	490.83	112.92	-0.17	0.11	150	850
2007	Science	83415	490.88	112.82	-0.19	0.10	150	850

Year	Subject	N	Mean	S.D.	Kurtosis	Skew	Min	Max
2008	Science	88674	493.49	112.71	-0.14	0.08	150	850
2009	Science	103713	497.13	113.20	-0.22	0.07	150	850
2010	Science	111233	499.11	112.30	-0.23	0.05	186	850
2011	Science	100420	493.76	112.43	-0.11	0.11	150	850
2004	History and Social Sciences	91430	493.01	106.89	-0.36	0.54	227	835
2005	History and Social Sciences	88536	491.17	110.96	-0.17	0.09	150	850
2006	History and Social Sciences	91114	490.66	110.93	-0.18	0.10	150	850
2007	History and Social Sciences	110313	490.49	110.99	-0.15	0.09	150	850
2008	History and Social Sciences	109408	492.62	110.99	-0.15	0.08	162	850
2009	History and Social Sciences	120903	494.58	111.91	-0.15	0.08	150	850
2010	History and Social Sciences	129512	496.93	111.43	-0.13	0.06	150	850
2011	History and Social Sciences	120566	492.93	110.94	-0.08	0.10	150	850
2004	NEM	127022	553.24	101.34	-0.64	0.17	208	826
2005	NEM	125053	557.33	101.19	-0.65	0.16	229	826
2006	NEM	129548	553.89	100.89	-0.64	0.19	208	826
2007	NEM	157971	544.36	101.69	-0.60	0.26	213	826
2008	NEM	160641	542.50	101.35	-0.58	0.27	208	826
2009	NEM	181835	539.52	102.55	-0.57	0.28	208	826
2010	NEM	196296	536.39	101.30	-0.56	0.31	208	826
2011	NEM	185994	534.51	101.22	-0.52	0.35	208	826
2004	Language and Mathematics	128705	490.81	107.95	-0.43	0.25	159.5	829.5
2005	Language and Mathematics	128284	490.03	105.58	-0.34	0.21	197.5	840.0
2006	Language and Mathematics	132944	490.47	105.47	-0.33	0.26	169.5	850.0
2007	Language and Mathematics	163049	490.53	104.76	-0.36	0.26	177.0	838.0
2008	Language and Mathematics	164927	493.26	106.26	-0.37	0.20	160.5	850.0
2009	Language and Mathematics	185930	496.27	107.31	-0.44	0.18	175.5	845.5
2010	Language and Mathematics	200520	498.30	107.28	-0.40	0.17	185.5	846.0
2011	Language and Mathematics	187819	494.67	105.24	-0.30	0.27	177.5	843.0

Table 143: Descriptive Statistics for PSU Raw Scores by Year and Subtest

Year	Subject	N	Mean	S.D.	Kurtosis	Skew	Min	Max
2004	Language and Communication	128700	37.16	13.81	-0.65	0.19	1	78
2005	Language and Communication	128277	36.47	14.67	-0.75	0.22	1	78
2006	Language and Communication	132940	35.75	15.20	-0.66	0.35	1	79
2007	Language and Communication	163038	34.37	15.28	-0.58	0.46	1	79
2008	Language and Communication	164904	34.09	16.25	-0.71	0.43	1	79
2009	Language and Communication	185914	36.67	16.51	-0.79	0.36	1	80
2010	Language and Communication	200490	35.76	16.20	-0.84	0.29	1	79
2011	Language and Communication	187801	35.38	14.77	-0.64	0.34	1	78
2004	Mathematics	128638	25.64	17.04	-0.35	0.83	1	70
2005	Mathematics	128165	25.71	17.59	-0.44	0.80	1	70
2006	Mathematics	132668	21.60	16.76	0.27	1.11	1	70
2007	Mathematics	162697	21.57	16.55	0.30	1.12	1	70
2008	Mathematics	164707	22.67	16.75	0.09	1.04	1	70
2009	Mathematics	185588	22.71	17.40	-0.07	1.01	1	72
2010	Mathematics	200078	23.24	17.27	0.01	1.01	1	70
2011	Mathematics	187386	21.53	17.04	0.53	1.23	1	70
2004	Science	67486	25.41	13.96	1.07	1.21	1	80
2005	Science	68128	26.69	14.93	0.67	1.13	1	78
2006	Science	72080	25.85	15.52	0.71	1.12	1	80
2007	Science	83392	26.79	16.10	0.34	1.01	1	79
2008	Science	88617	26.26	17.26	0.35	1.05	1	80
2009	Science	103643	27.14	16.81	0.33	1.03	1	80
2010	Science	111121	24.73	16.67	0.58	1.12	1	80
2011	Science	100313	23.43	16.98	0.92	1.26	1	80
2004	History and Social Sciences	91427	30.66	13.62	-0.25	0.57	1	75
2005	History and Social Sciences	88524	31.87	14.49	-0.52	0.44	1	73
2006	History and Social Sciences	91103	31.78	15.02	-0.53	0.47	1	75
2007	History and Social Sciences	110285	31.21	15.67	-0.57	0.47	1	75
2008	History and Social Sciences	109361	30.62	15.98	-0.46	0.57	1	75
2009	History and Social Sciences	120866	31.19	16.18	-0.49	0.55	1	75
2010	History and Social Sciences	129462	30.38	16.31	-0.43	0.62	1	75
2011	History and Social Sciences	120505	28.42	16.11	-0.30	0.70	1	75
2004	Language and Mathematics	128705	62.78	28.49	-0.49	0.58	2	145

Year	Subject	N	Mean	S.D.	Kurtosis	Skew	Min	Max
2005	Language and Mathematics	128283	62.15	30.20	-0.60	0.57	1	146
2006	Language and Mathematics	132944	57.30	29.86	-0.21	0.78	1	149
2007	Language and Mathematics	163049	55.89	29.88	-0.15	0.84	1	147
2008	Language and Mathematics	164919	56.73	31.08	-0.35	0.75	1	148
2009	Language and Mathematics	185920	59.33	31.99	-0.45	0.72	1	149
2010	Language and Mathematics	200512	58.95	31.56	-0.48	0.68	1	148
2011	Language and Mathematics	187815	56.86	29.78	-0.08	0.86	1	147

The first phase of the analysis dealt with examining trends in PSU scores over time and identifying variables that moderated those trends. Two approaches were taken to examine potential differences in trends. The first employed a factorial analysis of variance (ANOVA) with years and subpopulations (i.e. gender, school type, curricular branch) as the grouping variables. The results of those analyses are presented in Table 144 for Language and Communication and Mathematics combined scale score and in Appendix G for scale scores of individual subtests. Due to the very large sample sizes associated with these analyses, even very small differences between groups will show up as significant. Accordingly, we interpreted measure of effect (Cohen's f) in addition to statistical significance. To interpret the effect size results, we used Cohen's (1992) rules of thumb for magnitude of effect:

f of .10 = "small"
f of .25 = "medium"
f of .40 = "large"

In addition to the analysis of variance, the evaluation team plotted trend lines disaggregated by subpopulation for PSU scores across years. The 95% confidence intervals were plotted around the trend lines to facilitate the comparisons of subgroups. The confidence intervals generated are typically quite narrow as the samples used to generate them were large. The results for the Mathematics and reading combined scale score are presented in Figure 44 through Figure 49. The trend lines by subgroup for all tests are presented in Appendix H. Summary data for subpopulations across years are presented as tables in Appendix I. Statistical analyses and trend lines for raw scores are presented in Appendix K and Appendix L.

Table 144: Factorial ANOVAs of PSU Language & Mathematics Combined by Year, Gender, Type, Region and Curricular Branch

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Language & Mathematics	Year	8	13897780	154.93	0.00	0.03
Language & Mathematics	Gender	1	104436404	9313.73	0.00	0.08
Language & Mathematics	Year*Gender	8	1656907	18.47	0.00	0.01
Language & Mathematics	Error	1457312	16341079856			
Language &	Year	8	12595725	140.01	0.00	0.03

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Mathematics Language & Mathematics	Region	2	53611219	2383.64	0.00	0.06
Language & Mathematics	Year*Region	16	6154387	34.20	0.00	0.02
Language & Mathematics	Error	1452877	16338554404			
Language & Mathematics	Year	8	12247336	161.11	0.00	0.03
Language & Mathematics	Type	3	2577337253	90413.64	0.00	**0.43**
Language & Mathematics	Year*Type	16	23183634	152.49	0.00	0.04
Language & Mathematics	Error	1443266	13713945642			
Language & Mathematics	Year	8	13893682	191.64	0.00	0.03
Language & Mathematics	Branch	8	3170946775	43737.84	0.00	**0.49**
Language & Mathematics	Year*Branch	56	70007163	137.95	0.00	0.07
Language & Mathematics	Error	1457243	13206071541			
Language & Mathematics	Year	8	9711815	134.62	0.00	0.03
Language & Mathematics	SES Quintile	4	1076255539	29835.93	0.00	**0.31**
Language & Mathematics	Year*SES	32	63502937	220.05	0.00	0.07
Language & Mathematics	Error	1247179	11247206986			

(Note: Medium and large effect sizes are shown in bold face.)

The results illustrate that differences between subpopulations were almost always statistically significant. This is not surprising. Given the substantial sample size and ample statistical power associated with these tests, even small differences would be revealed as significant. Accordingly, it is more appropriate to interpret the effect sizes associated with the tests of significance and relationships indicated by the trend lines.

An examination of the overall trend lines not broken down by subgroup suggest the scores were fairly consistent from 2004 to 2006 and from 2006 to 2007, a crossing point from the two trend lines we investigated (2004-2006 and 2007-2011). Starting in 2007, the scores steadily increase until 2010; then there is a slight drop-off from 2010 to 2011. An examination of the trend lines by subgroup suggest that this increase in scores is due largely to an increase of high scores from private schools and schools with a Scientific-Humanistic curricular branch, while the scores at subsidized schools, public schools, and schools with a Technical-Professional curricular branch stayed the same or even decreased. As an example, Appendix I indicates the mean PSU Mathematics scale score for private schools in 2004 was 581 and 621 in 2011 – an increase of approximately 40 points or about half of a standard deviation. Conversely, the mean PSU Mathematics score for municipal schools in 2004 was 465 and in 2011 it was 464 – a decrease of one scale score point.

Additionally, the effects for school type and curricular branch were substantial. The average effect for school type was approximately 0.43 and the average effect across subtests for curricular branch was approximately 0.49—large by Cohen's rule of thumb. Students in Scientific-Humanistic tracks scored substantially better than student in Technical-Professional tracks, and students in private schools perform substantially better than those in subsidized and municipal schools. This substantial difference appeared in 2004 and increased since then until 2011.

There was also a significant effect for gender and region, but the effect was smaller than for curricular branch and school type. Appendix G and Appendix H indicate that males performed better than females across all years on the Science, Mathematics and History subtests (though effect sizes ranging from .12 to .14 were small).

The gender difference was smallest for the Language subtest (f=.01). Females actually had higher mean scores in Language in 2004 and 2011. The negligible effect size for gender*year (f= .01 to .03) indicates the effect for gender within subtests is fairly consistent across years.

There is a consistent and moderate effect for SES of approximately .28 across subtests. The finding that SES is positively and significantly related to PSU test scores is consistent with previous research conducted in the United States (Sirin, 2005; Sackett, et al., 2007; Zwick & Greif-Green, 2007), indicating that the phenomenon is also present in Chile. This effect is illustrated in the trend lines and the tables for PSU score by SES quintile in Appendix H and Appendix I. The trend lines plot a mean PSU score that is higher for each respective quintile. Moreover, there are no instances of crossover, i.e., the trend line for quintile B is always higher than the trend line for quintile A and always lower than the trend line for quintile C, etc.

Region was also a significant moderator of PSU scores. Students in the central region consistently outperformed students in the northern and southern regions of the country. Students in the southern region usually performed better than students in the northern region, but the gap closed somewhat in the 2011 administration. Despite performing the best on all subtests of the PSU, students in the central region tended to have significantly lower NEM scores than their counterparts in the northern and southern regions.

Next, the evaluation team sought to determine the relationship these variables have with each other. Given there is a moderate relationship between SES and PSU scores, the evaluation team attempted to determine if covariation between SES and region may explain some of the differences in PSU scores among regions. The mean SES quintile for each region is presented below in Table 145.

Table 145: Mean Socioeconomic Status by Region

Region	N	Mean SES Quintile	S.D.	Minimum	Maximum
Central	649489	3.13	1.40	1	5
North	146321	3.15	1.39	1	5
South	447510	2.80	1.38	1	5

The results indicate that SES does not explain the superior performance of students in the central region. The northern region had the highest mean SES, but typically had the lowest PSU scores. The evaluation team then sought to find if the distribution of curricular branch

across regions may explain the differences in scores among regions. Those results are presented in Table 146.

Table 146: Percentages of Curricular Branch by Region

Curricular Branch	Region		
	North	Central	South
Scientific-Humanistic	12%	53%	35%
Technical-Professional	11%	54%	35%

Table 146 indicates the distribution of Scientific-Humanistic and Technical-Professional schools are approximately equal across regions and cannot explain the difference in scores between regions. The breakdown of type of school by region is presented in Table 147. These results indicate that the central region has a higher percentage of private schools than the southern and northern regions. Given the large effect in favor of private schools (.35 to .43), this result may help explain the high performance of students in the central region when compared to the northern and southern regions.

Table 147: Percentages of School Type by Region

Region	Type of School		
	Private	Subsidized	Municipal
North	6%	47%	47%
Central	15%	53%	32%
South	7%	39%	53%

Lastly, the evaluation team looked at the relationship between curricular branch and school type. The frequency of curricular branch by school type is presented below in Table 148.

Table 148: Percentages of School Type by Curricular Branch

Percentages of Curricular Branch by School Type			
Curricular Branch	Type of School		
	Private	Subsidized	Municipal
Scientific-Humanistic	16%	48%	36%
Technical-Professional	0%	45%	55%

There were virtually no schools in the Technical-Professional branch, while 16% of Scientific-Humanistic schools were private. There is the possibility that school type is affecting the results for curricular branch, or vice-versa. To address this question, an analysis of variance that allowed us to look at the separate effects of curricular branch, school type, and any interaction was conducted. Those results are presented in Table 149 and Table 150.

Table 149: ANOVA of School Type by Curricular Branch

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Language & Mathematics	Curricular Branch	2	1669378037	94088.49	0.00	0.36
Language & Mathematics	School Type	3	2565158516	96383.95	0.00	0.45
Language & Mathematics	Ed. Branch*School Type	4	765421082	21570.1	0.00	0.25
Language & Mathematics	ERROR	1436961	12747739952			

The results indicate that while there is substantial relationship between curricular branch and school type as indicated by the effect size for the interaction (f=.25), there are also strong effects for curricular branch and school type individually as indicated by the effect sizes associated with those tests (.36 and .45). The means for each cell are presented below in Table 150.

Table 150: Mean Values for Language and Math Combined by School Type and Curricular Branch

Curricular Branch		N	Mean	S.D.
Scientific-Humanistic	Private	163122	603.62	98.79
	Subsidized	484475	512.40	96.26
	Municipal	361166	478.70	103.99
Technical-Professional	Private	91	493.93	67.19
	Subsidized	194333	442.56	79.70
	Municipal	233394	438.85	79.73

The trend lines for the overall Language and Communication and Mathematics tests and by subgroup are presented in Figure 44 through Figure 49 on the following pages. The trend lines for each individual subtest are located in Appendix H.

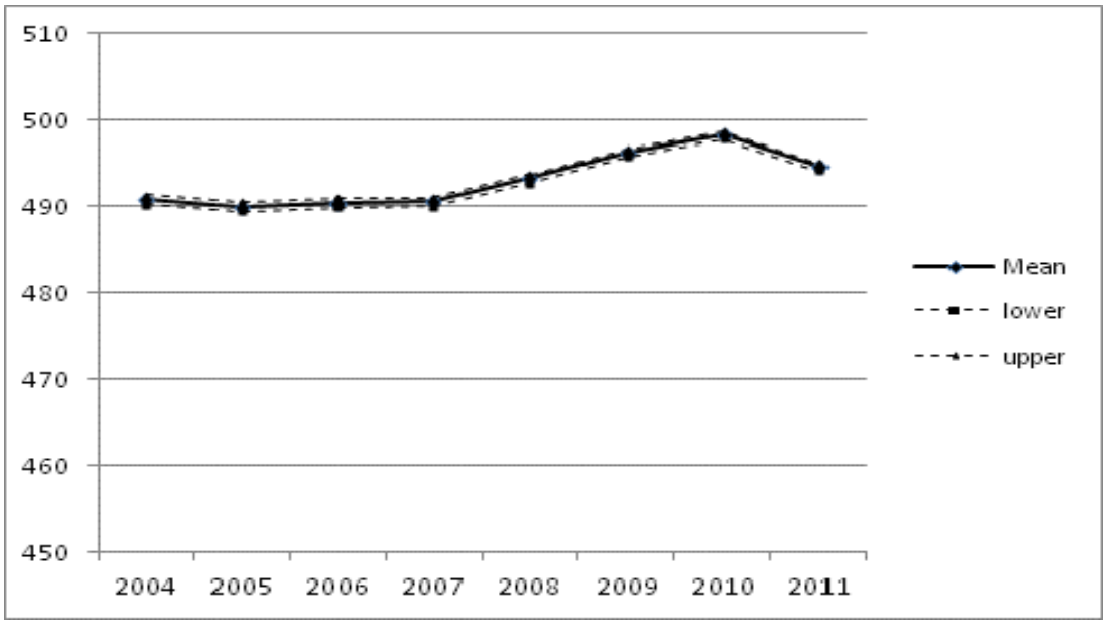


Figure 44: Trend Analysis of PSU Language and Mathematics Combined Score

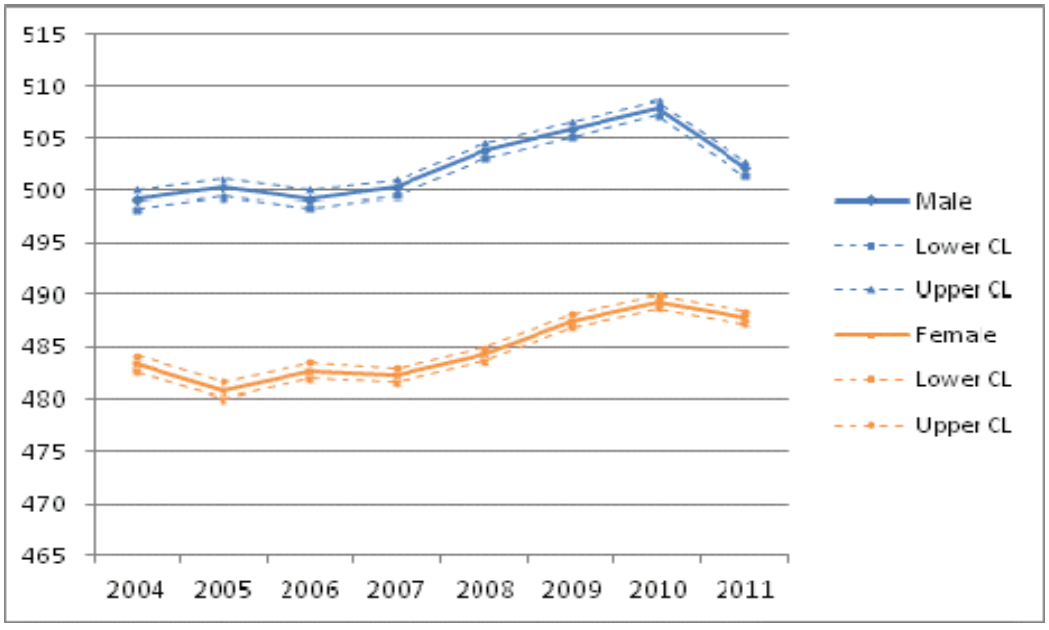


Figure 45: Trend Analysis of PSU Language & Mathematics by Gender

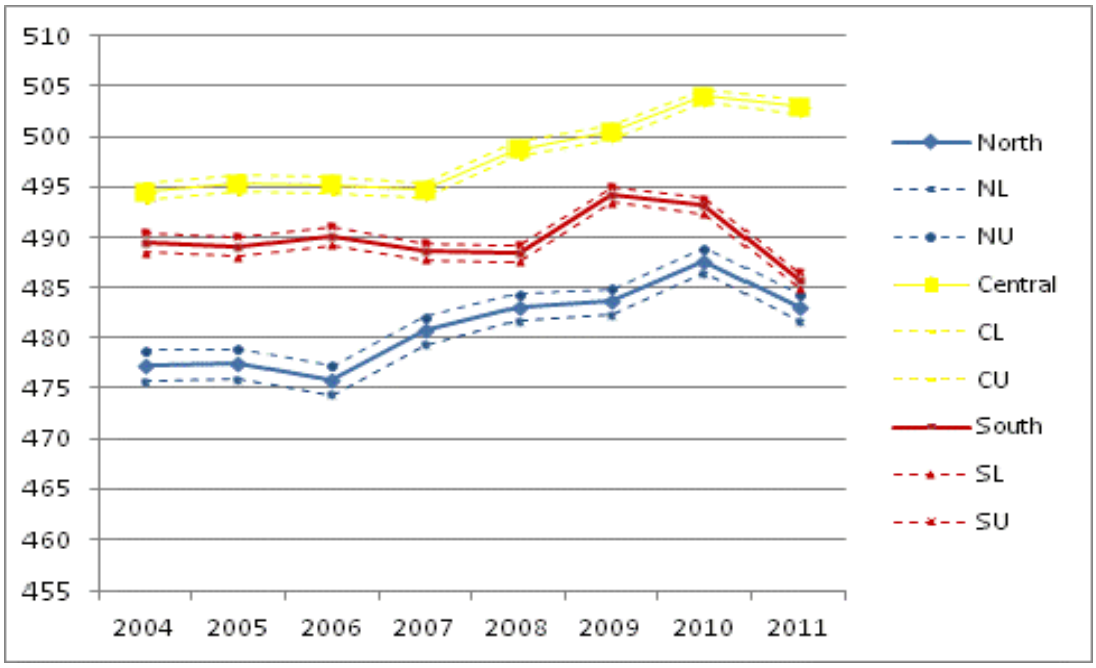


Figure 46: Trend Analysis of PSU Language & Mathematics by Region

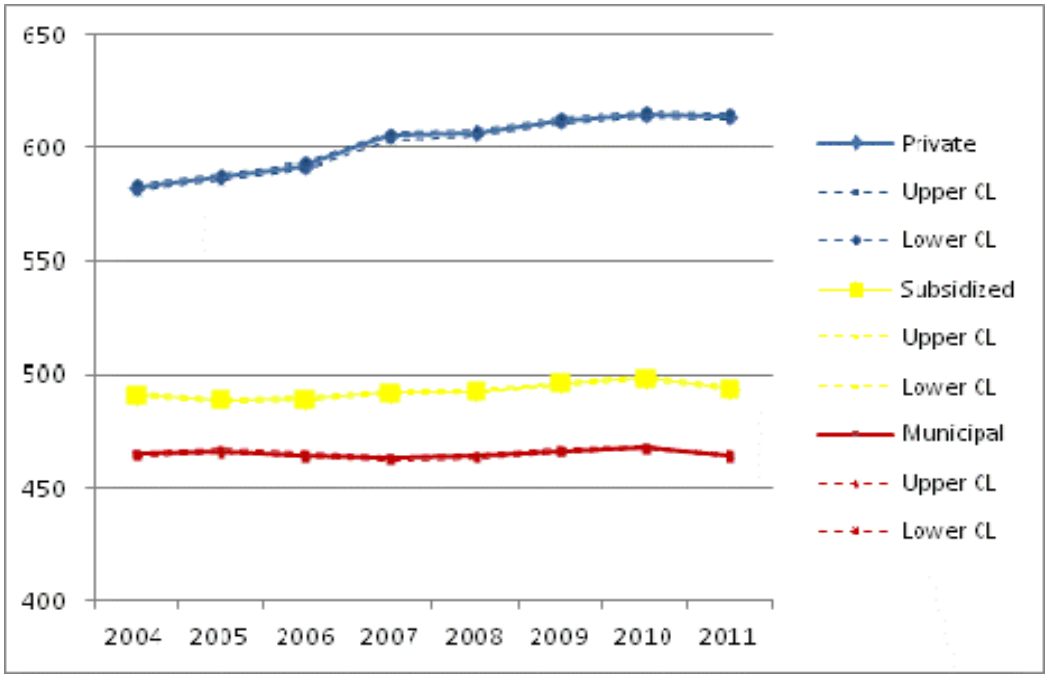


Figure 47: Trend Analysis of PSU Language & Mathematics by School Type

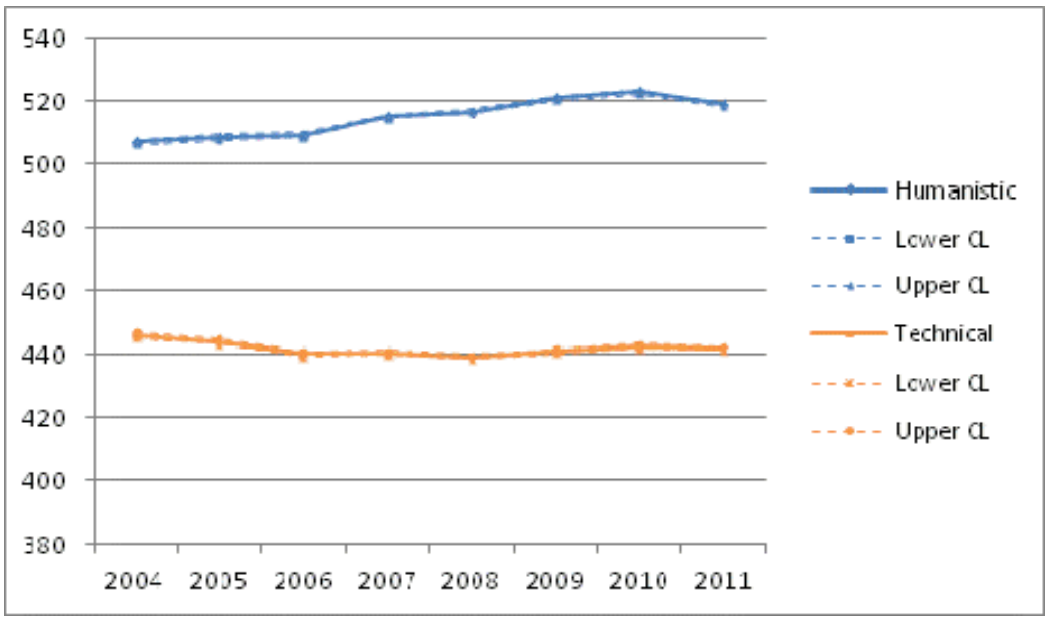


Figure 48: Trend Analysis of PSU Language & Mathematics by Curriculum

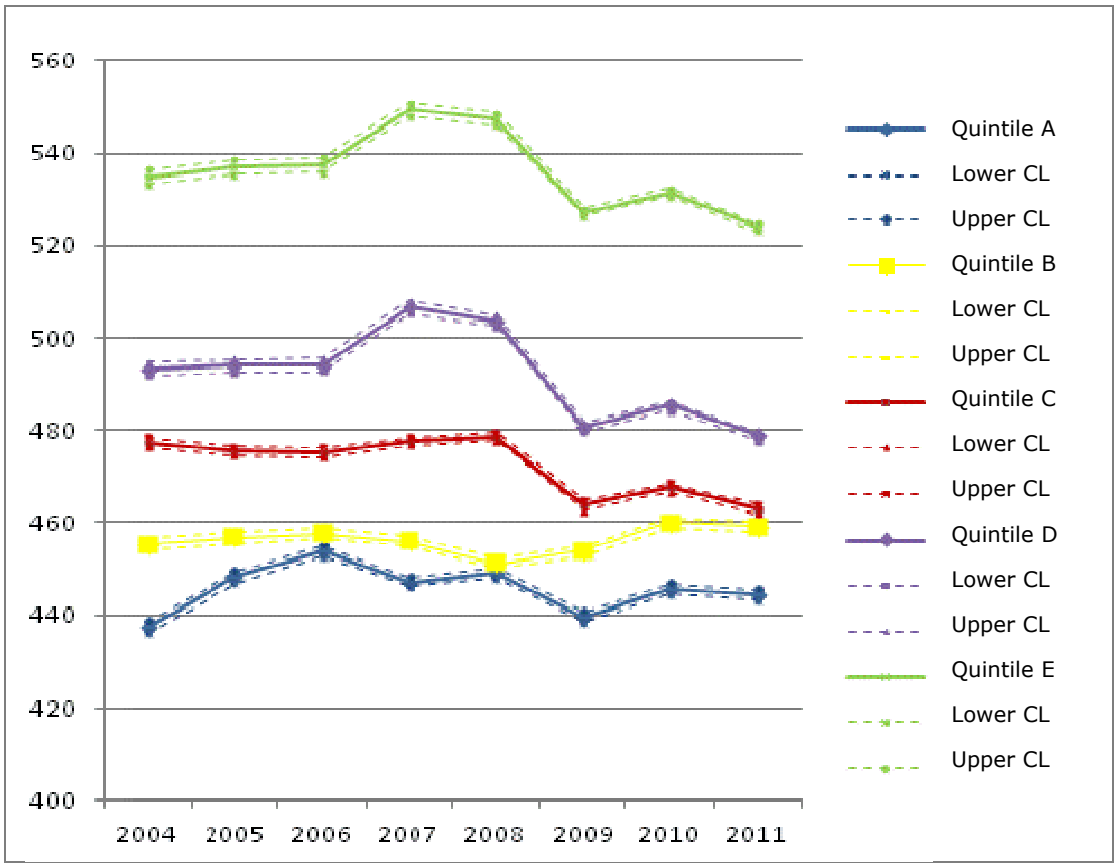


Figure 49: Trend Analysis of PSU Language & Mathematics by SES Quintile

Part II:

The second phase of the cohort analysis sought to examine the relationship between NEM and PSU scale scores within each high school. Additionally, the evaluation team examined the effect of school-level characteristics on the covariation between NEM and PSU test scores.

In the first stage, the relationship between high school performance (NEM) and PSU scale scores was examined using Pearson product moment correlations within each high school. These correlations were aggregated into a sample weighted correlation using the procedure outlined by Hunter and Schmidt (1990). Those results are presented below in Table 151.

Table 151: Correlation Analyses

Year	PSU Subtest	Weighted r	Year	PSU Subtest	Weighted r
2004	Language and Communication	0.35	2004	History and Social Sciences	0.25
2005	Language and Communication	0.44	2005	History and Social Sciences	0.32
2006	Language and Communication	0.48	2006	History and Social Sciences	0.33
2007	Language and Communication	0.47	2007	History and Social Sciences	0.33
2008	Language and Communication	0.48	2008	History and Social Sciences	0.35
2009	Language and Communication	0.47	2009	History and Social Sciences	0.33
2010	Language and Communication	0.47	2010	History and Social Sciences	0.32
2011	Language and Communication	0.42	2011	History and Social Sciences	0.32
2004	Mathematics	0.44	2004	Science	0.46
2005	Mathematics	0.50	2005	Science	0.55
2006	Mathematics	0.53	2006	Science	0.56
2007	Mathematics	0.54	2007	Science	0.56
2008	Mathematics	0.54	2008	Science	0.58
2009	Mathematics	0.55	2009	Science	0.56
2010	Mathematics	0.52	2010	Science	0.57
2011	Mathematics	0.51	2011	Science	0.57

The results indicate there is a significant relationship between NEM and PSU scale scores across all subtests. Additionally, the relationship appears strongest for Science and Mathematics, where the weighted correlations are between .50 and .60. The weighted correlations for NEM and Language and Communication are typically between .45 and .50 and the correlations for History and Social Sciences and NEM are typically between .30 and .35. Interestingly, the values are lower for 2004 than the other years. The pattern may be the result of decisions made in subsequent years to increase the alignment of PSU test frameworks to Chile's national curriculum. The results for the raw score analysis is presented in 0.

In the second phase of this analysis, a hierarchical linear modeling approach was implemented in order to capture the degree to which school-level characteristics influenced the relationship between NEM and PSU scores. Main effects and interactions were modeled and interpreted. The model fitted for each subtest is presented below:

$$Y_{ij} = Y_{00} + Y_{10}(\text{NEM}) + Y_{01}(\text{SchoolSES}) + Y_{02}(\%\text{Female}) + Y_{03}(\text{Region}) + Y_{04}(\text{Type}) + Y_{05}(\text{Curr. Branch}) + Y_{11}(\text{SchoolSES})*(\text{NEM}) + Y_{11}(\text{NEM})*(\text{SchoolSES}) + Y_{11}(\text{NEM})*(\%\text{Female}) + Y_{13}(\text{NEM})*(\text{Region}) + Y_{14}(\text{NEM})*(\text{Type}) + Y_{15}(\text{NEM})*(\text{Branch}) + u_{oj} + r_{ij}$$

Y_{00} represents the grand mean. The level-1 predictor is NEM(Y_{10}). The level-2 predictor are school-level SES(Y_{01}), % of students that are female (Y_{02}), region (Y_{03}), school type (Y_4), and curricular branch(Y_{05}). The cross-level interaction terms are Y_{11}(NEM)*(SchoolSES), Y_{12}(NEM)*(%Female), Y_{13}(NEM)*(Region), Y_{14}(NEM)*(Type), and Y_{15}(NEM)*(Branch). The term u_{oj} represents the random variation of schools around the intercept and the term r_{ij} represents the random variation of students within schools.

The results for Language and Mathematics combined scale score across all years are presented below in Table 152. The analyses for scale score across years by all subtests are presented in Appendix A. The same general trends hold across subtests. The slope is positive for NEM centered. These results align with the results of the previous analysis in that the slopes are steeper for Science and Mathematics than for Language and History. The analyses for raw scores are presented in Appendix N.

Table 152: Results of Hierarchical Linear Modeling Analysis – Language and Math Combined

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		545.62	1.55	2414	352.81	0.00
NEM		0.14	0.00	231149	34.59	0.00
School SES		2.87	0.08	231149	35.86	0.00
% Female		-7.84	2.46	231149	-3.19	0.00
Region	Central	12.57	0.80	231149	15.71	0.00
	North	-17.67	0.97	231149	-18.24	0.00
	South	0.00				
School Type	Private	63.81	1.09	231149	58.41	0.00
	Subsidized	19.59	0.82	231149	23.94	0.00
	Municipal	0.00				
Curricular Branch	Scientific	21.56	0.66	231149	32.53	0.00
	Technical	0.00				
SES*NEM		0.01	0.00	231149	7.37	0.00
NEM*%Female		0.03	0.00	231149	5.37	0.00
NEM*Region	Central	-0.03	0.00	231149	-12.87	0.00
	North	-0.03	0.00	231149	-6.97	0.00
	South	0.00				
NEM*Type	Private	0.15	0.00	231149	39.56	0.00
	Subsidized	0.04	0.00	231149	14.23	0.00
	Municipal	0.00				
NEM*Branch	Scientific	0.19	0.00	231149	56.11	0.00
	Technical	0.00				

For Language and Mathematics combined, the estimate of the grand mean (Y_{00}) is 545.62. In other words, the school mean Language/Mathematics achievement is 545.62, when all other predictors are controlled for. The negative value for % female indicates that the higher the percentage of female students at the school, the lower the mean PSU score is. This relationship holds for all subtests except for the Language subtest, where the mean PSU score for a school increases as the percentage of females at the school increases. This finding aligns with the trend analysis that indicated there is not as much of a gender gap on the Language test as there is for the other subtests.

The main effect for the categorical variables indicate the intercepts for the models are significantly different depending on which classification group the student falls in. For example, the intercept for private schools is 63.81 higher than for municipal schools when all other predictors are controlled for. Of the continuous predictors, school-level SES and NEM are positively related to the PSU score and percentage of female students in the school is negatively related to the PSU score.

The research question of interest dealt with identifying school-level characteristics that moderated the relationship between NEM and PSU test scores. To examine this, the results of the interaction of terms are of particular interest. The significant interaction for SES and NEM indicates the slope for NEM differs depending on the mean SES of the school. The positive value of the interaction term indicates, for each one unit increase in school-level SES, the slope increases by .01.

The significant interactions for NEM and the categorical predictors indicate the slope for NEM differs for the respective groups. Most notably, there is a stronger relationship between NEM and PSU score in private schools and schools that have a Scientific-Humanistic curricular branch. To illustrate, a fitted model for a student in a subsidized school providing the Scientific-Humanistic curricular branch in the northern region is presented below.

PSU = 586.77 +.34NEM + 2.87SES – 7.84%Female + .01SES*NEM + .03%Female*NEM

The slope for NEM here is larger than in Table 152 (.34 vs. .14) as the above model is for schools providing the Scientific-Humanistic curricular branch, which have a steeper slope than the schools providing the Technical-Professional curricular branch when all other variables are controlled for. In the same vein, the slope is steeper for private schools than for subsidized and municipal schools when all other variables are controlled for. In other words, not only is the performance of students in private schools and schools with a Scientific-Humanistic curricular branch higher, but the relationship between NEM and PSU scores is stronger as well. Also, while the percentage of females at a school tends to impact negatively the intercept, it tends to increases the slope for NEM.

EVALUATION

The primary focus of this research was to analyze trajectories of PSU scores from 2004 through 2011 and identify variables that moderated those trajectories. Additionally, differences between generations of the PSU were examined – 2004 through 2006 and 2007 through 2011.

Subtest scores from 2004 through 2006 stayed relatively stable from year to year. Scores did not change that much from 2006 through 2007 indicating the change initiated that year did not affect PSU scores. Scores across subtests did steadily increase from 2007 through 2010. This result reiterates a need for test equating to take place on yearly basis.

The evaluation team would like recommend the need for carrying analyses of equating invariance as well. The trend analyses by subpopulations have shown differences in PSU scale scores. A closer examination of the trend-lines by subpopulations suggests this increase was driven by students at Private schools and schools with a Scientific-Humanistic curricular branch. Mean scores for those groups steadily increased from 2007 through 2010 while the trend lines for Municipal schools, Subsidized schools, and schools with a Technical-Professional curricular branch indicated little to no change from year to year or even decreased. The difference in scores between Private schools and Subsidized or Municipal schools and the difference between schools with a Scientific-Humanistic curricular branch and schools with a Technical-Professional curricular branch are substantial. Furthermore, these gaps have increased between the years of 2004 and 2011.[13]

These results are not particular to Chile since there is international evidence that shows applicants from private high schools outperforming applicants from public high schools on university admissions tests. For example, in the U.S. the 2009 College-Bound Seniors report documented that a typical applicant from the private high schools outperforms a typical applicant from the public high schools on SAT scores on critical reading and mathematics by 54 and 68 scale points, respectively (College Board, 2009). Moreover, these differences in performance have remained relatively stable over multiple administration years (College Board 2010, 2011).[14]

On all subtests other than Language and Communication, female students performed significantly lower than Male students, despite having higher high school NEM. The gap was much closer for Language and Communication, with females even outperforming males for a couple of years (see Figure 56 through Figure 60) available from the appendices of the report). Disparity on university admissions test scores between male and female applicants is a pervasive finding in university admissions testing internationally. In the U.S., for example, a typical male applicant attained larger scores than a typical female applicant on the SAT critical reading and mathematics tests (College Board 2009, 2010, 2011). For the 2009 college-bound total population, the gender gap for SAT critical reading (5 scale score

[13] Because of Chile's curriculum reform, there has been a gradual implementation of the new curriculum in the content of the PSU tests. This began in 2003 and was completed in 2005. Strictly speaking, this change in the content of the PSU may have affected these trends.

[14] The evaluation team was unaware of an international study in which longitudinal data (i.e., panel study) have been analyzed under the two modalities (curriculum and aptitude) and disaggregated by subpopulations relevant to Chile (e.g., Scientific-Humanistic versus Technical-Professional curricular branches). What we have done is to compare the PSU results by subgroup to another widely used measure for university admissions, viz., the SAT. To this end, the evaluation team would have preferred relying on a panel study having applicants taking both curriculum-based and aptitude-based admissions tests.

points) is about six times smaller than the gender gap for SAT mathematics (35 scale score points). Interestingly, the gender gap has remained relatively stable over multiple years of test administrations.

Students in the central region performed significantly better than students in the southern and northern regions. A more thorough examination suggested the reason for this difference may be due to a larger percentage of private schools in the central region than in the northern and southern regions (see Table 147). The role that socio-economic status (SES) plays on university admissions test scores for college bound students has been documented internationally. For example, in the U.S. across multiple indicators of college-bound seniors' SES (e.g., plans to apply for financial aid, family income, and highest level of parental education), average SAT critical reading and mathematics scores are consistently larger for the higher SES groups (College Board 2009, 2010, 2011).

A secondary goal was to explore the relationship between high school performance (NEM) and PSU scores and identify characteristics of high schools that affected the covariation of those two variables. There was a significant weighted correlation between NEM and PSU scores, more so for the Science and Mathematics subtests than for the Language and Communication and History and Social Sciences subtests.

We followed this basic analysis by exploring the extent to which school-level characteristics moderated the relationship between NEM and PSU scores using the statistical technique of hierarchical linear modeling. The results indicated that school-level characteristics moderate the slope of the regression line. The variables with the largest effect on the slope were school type and curricular branch. The slope for private schools was anywhere from .14 to .22 steeper than for municipal schools, depending on the test subject for the model fitted and anywhere from .12 to .26 higher for schools offering the Scientific-Humanistic curricular branch than for those offering the Technical-Professional curricular branch. Not only do private schools and schools with a Scientific-Humanistic orientation outperform public and technical schools on the PSU, the relationship between NEM and PSU scores is stronger in these schools. While the mean scores for girls were typically lower, the slope typically increased anywhere from .03 to .06 for every one unit increase in the percentage of females at the school. There was also a slight increase in the slope as the SES increased.

In summary, the differences in PSU test scores over several years of administration have shown fluctuations that indicate a strong need for equating to take place within Chile's university admissions testing program. Nevertheless, within-year comparisons are still possible, and they show important differences in PSU tests scores for relevant subpopulations. The patterns of PSU test scores across relevant demographic variables are akin to patterns observed internationally with college-bound seniors. The gap size, on the other hand, is large for the Chilean population of college bound students, particularly for the subpopulations based on type of school and socio-economic status.

RECOMMENDATIONS

1. The evaluation team recommends carrying out test score equating on a yearly basis. Along this line, we recommend careful inspection of comparability PSU test scores between years. In Chile PSU test scores can be used for two consecutive admission processes. After accounting for lack of test score equating and changes on PSU test specifications (e.g., mathematics test increased length in the 2012 admission process), equity of PSU test score between adjacent years is at stake. It should be a matter of indifference for applicants whether they take the PSU test in 2011 or in 2012.

2. The evaluation team recommends inspecting invariance of equating functions across relevant subpopulations of applicants. These types of verifications are germane to developing validity evidence on meaning of test scores. Strong equating results should be invariant across subpopulations; otherwise, linking studies to align score scales should be performed to allow for comparisons of PSU test scores.

BIBLIOGRAPHY

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kreft, I., & de Leeuw, J. (1998*). Introducing multilevel modeling*. Sage Publications Ltd.

Raudenbush, S. W., & Bryk, S. A. (2002). *Hierarchical linear models: applications and data analysis methods.* Thousand Oaks, CA: Sage Publications

Sackett, P., Kuncel, N., Arneson, J., Cooper, S., & Waters, S. (2009). *Socioeconomic status and the relationship between the SAT and freshman GPA: An analysis of data from 41 colleges and universities* (Research Report No. 2009-1). The College Board.

Sirin, S. R. (2005) Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*, 3, 417-543.

White, K. R. (1983) The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *9*, 3, 461-481.

Zwick, R., & Greif-Green, J. (2007). Perspectives on the correlation of scholastic assessment test scores, high school grades, and socioeconomic factors. *Journal of Educational Measurement*, *44*, 1, 1-23.

Objective 2.4. PSU predictive validity: To complement predictive validity on population groups throughout administration years, considering the differences experienced in those taking the PSU and the test variations since its implementation (2004), which shall contemplate a differential validity analysis and possible differential prediction of the PSU through year and type of career, considering subgroups defined by gender, dependence and education mode

ABSTRACT

The purpose of this study was three-fold: (1) document the ability of PSU test scores and high school academic performance (NEM and high school ranking) to predict university students' academic outcomes; (2) document incremental prediction value of the variable ranking; and (3) examine the extent to which PSU test scores and high school academic performance exhibits differential prediction for relevant demographic variables (Gender, SES, Region, High School Curricular Branch, Type of School). The study relied on three university academic outcomes: (1) first year grade point average, (2) second year grade point average, and (3) university completion. The study relied on longitudinal data sets for university admissions that spanned 2004-2012. DEMRE provided the databases with PSU test scores and high school grade point average (NEM), and MINEDUC provided the databases with students' university academic outcomes and their high-school ranking score. Linear and logistic regression analyses were run separately for each career within a university and summarized across careers and universities. Corrections for restriction of range, involving variances and standard deviations of PSU test scores from the population of university-bound seniors (i.e., population of university applicants), were applied to the Pearson validity coefficients from the population of university students (Gulliksen, 1987). PSU predictive validity results were weighted so as to assign more weight to larger sample sizes (Hunter & Schmidt, 1990). Incremental prediction validity of the variable ranking was computed by fitting base and revised models to the data set. The revised model used ranking as an additional predictor. The effect of the variable ranking was documented by computing the difference in variance reduction of university outcomes (e.g., the revised model minus the based model). Analyses of differential validity were carried out by demographic variables with estimates of standardized residuals computed within careers and disaggregated by demographics. For summative purposes the individual students' residuals were averaged across careers and admission years before their disaggregation by demographic variables[15].

Results showed that PSU test scores predict university students' academic outcomes, particularly first and second year grade-point average. Although PSU test scores showed patterns of prediction that were similar to those from international studies, the magnitudes of the PSU prediction coefficients were found smaller than those observed internationally (Matter, et al., 2008). The variable ranking contributed to the reduction of uncertainty on students' academic outcomes after controlling for students' PSU test scores and students' NEM. The largest amount of variance reduction (7%) happened for students' university completion. Predictive and incremental validity results by type of career showed similar prediction patterns to those from the overall analyses. For example, PSU Mathematics and Science scores and high school academic performance (NEM and ranking) showed larger predictive capacity than PSU Language and Communication and History and Social Sciences scores. In addition, applicants' high school ranking showed incremental predictive validity

[15] The study also carried out predictive and incremental validity and prediction biases analyses by type of career, separately. MINEDUC provided the list of the type of career included in the analyses.

of university outcomes (over and beyond PSU test scores and NEM), although its contribution was smaller than the one found from the overall analyses.

Finally, prediction bias findings showed under-prediction patterns for females similar to those reported internationally (Matter, et al., 2008). Interestingly, the magnitude of the prediction bias was smaller than those reported internationally. Graduates from Technical-Professional high school buildings showed under prediction for short term university success outcomes; although the size was, for practical purposes, small. Other relevant demographic variables considered in the study resulted in negligible differential prediction bias. The PSU test scores and high school performance measures appear to result in comparable amounts of differential prediction validity for major demographic variables.

When examining predictive and incremental validity results by type of career, we saw similar prediction patterns to those from the overall analyses. For example, PSU Mathematics and Science scores and high school academic performance (NEM and ranking) showed larger predictive capacity than PSU Language and Communication and History and Social Sciences scores. In addition, applicants' high school ranking showed incremental predictive validity of university outcomes (over and beyond PSU test scores and NEM), although its contribution was smaller than the one found from the overall analyses.

INTRODUCTION

Predictive validity refers to the ability of test scores to forecast performance on a relevant criterion (AERA, APA, NCME, 1999). University admissions decisions are complex and involve multiple measures among which university admissions tests scores and high school academic performance are often the focal variables (i.e., predictors) while university academic performance measures are the outcome variable of interest (i.e., criterion). Higher education institutions have involved in their selection processes institutional preferences –defining admissions criteria and weights in prediction of university outcomes. Examples of university academic performance are first year university grade point average and cumulative university grade point average. Internationally, a long lasting recommendation for university admissions criteria is to involve some sort of aptitude/academic tests and prior academic performance of applicants (e.g., high school grades). These sorts of information have been found to be good predictors of university outcomes (Whitney, 1993). The best prediction models often involve all of the above predictors. Background variables (e.g., parents' educations and career plans) often add little to the reduction of uncertainty of university academic performance (Roberts & Noble, 2004).

International standards of testing have recommended inspecting validity of inferences made from test scores. If it is asserted, for example, that a given university admissions test will result in scores that predict future academic performance in university, evidence in support of that assertion needs to be collected and presented (AERA, APA, NCME, 1999; International Test Commission, 2000). Documentation of the predictability of the elements that compose the admission criteria is often attained through predictive validity studies involving correlations between predictors and outcomes. That is, prediction validity refers to a finding where university admissions criteria adequately correlate with university outcomes.

International testing programs are not content only to document that test scores adequately predict university admissions outcomes. Testing programs also aim at documenting degrees of differential university outcomes for populations of interest. Differential prediction validity is another kind of institutional validity study that is directed

at investigating the degree of similarities and differences in predicted outcomes among relevant subpopulations. The goal for this sort of validity study is to document the invariance of prediction across subgroups (Linn, 1982; Young, 2001). In other words, a differential prediction study seeks to determine whether a common prediction equation yields significantly different results for different groups of examinees. Ideally, a good university admissions test should predict academic performance in university equally well among subpopulations. In reality, university admissions tests may show patterns of over/under prediction favoring a specific subpopulation over the others.

Chile's university admissions process takes into account PSU test scores and high school academic performance (NEM) and combines these two indicators using a set of weights decided by each career and university. DEMRE applies weights to the PSU test scores and NEM values when producing admission scores that later inform the rank ordering of applicants to careers and universities of their preferences. In Chile, predictive validity studies have been conducted by scrutinizing the degree of relationship between PSU test scores, high school grade point average, and first-year university grade point average (*Comité Técnico Asesor*, 2010). These studies have noted the Pearson product-moment correlation coefficients between PSU test scores and NEM and first-year university grade point average. In these analyses, the unit of analysis was the career with a minimum of ten valid records. Results were interpreted qualitatively by spotting large numerical differences on average correlations. Findings from the study showed that PSU Mathematics and PSU Science tests resulted in a larger predictive validity coefficient than did NEM. In contrast, PSU Language and History tests resulted with the lowest predictive validity (*Comité Técnico Asesor*, 2010).

In Chile, differential predictive validity studies have been performed with within-career, linear single-predictor regression models involving PSU scores for the mandatory tests (i.e., Mathematics and Language and Communication), high school grade point average, and university selection scores as the predictor variables, and cumulative grade point average for first year of university as the criterion variable (*Comité Técnico Asesor*, 2010). The study reported the presence of small amounts of differential predictive validity similar to those reported internationally. Internationally, past studies using multiple linear regression models have shown a small amount of differential prediction. For example, Bridgeman et al. (2000) studied differential predictive validity of SAT verbal, SAT Mathematics, and high school grade point average with a sample involving multiple campuses in the U.S; and Maxey and Sawyer (1981) studied differential predictive validity of ACT subtests and high school grades with more than 250 colleges and universities from the U.S. Young (2001) reported similar patterns with a meta-analysis of research findings from multiple studies published between years 1974 and 2000.

Goals and Research Questions

The purpose of this study was three-fold: (1) document the ability of PSU test scores and academic performance in high school (high school GPA and high school ranking) to predict university academic performance; (2) document incremental prediction value of the variable ranking; and (2) examine the extent to which PSU test scores and high school academic performance exhibits differential prediction across relevant demographic variables (gender, SES, region, high school curricular branch, type of school).

For predictive validity, the study sought to give answer to the following research questions:

- What is the predictive validity of PSU test scores and high school grade point average (NEM) on university first year grade point average, second year grade point average and university graduation?

- What is the incremental predictive validity of the variable ranking (measured as a proxy variable from NEM) over and beyond PSU test scores and NEM on university first year grade point average, second year grade point average, and university graduation?

Differential predictive validity is another kind of institutional validity study that is directed to investigate the degree of similarity and difference in predicted outcomes among relevant subpopulations. With respect to differential predictive validity, this study sought to answer the following question:

- What is the differential predictive validity of PSU test scores and high school grade point average on university first year grade point average, second year grade point average, and university graduation for the following variables:

    - Gender,
    - Socio-economic level,
    - Region,
    - High school curricular branch, and
    - Type of high school funding?

METHODOLOGY

Demographic Variables

The study relied on students' survey data collected as part of the registration process prior to taking PSU test battery. Applicants' self reported information on family income and parental education was obtained from DEMRE data sets and utilized to construct a socio-economic status index (SES). Specifically, SES was computed by using information from parental income and the father's and mother's educational level. The process to compute socio-economic variables is available in the PSU evaluation report. DEMRE also provided information on applicants' gender and school information. The following list shows applicants' background variables used in this study.

- Gender: Male or Female;
- Type of high school: Private, Subsidized or Municipal.
- Curricular branch: Scientific-Humanistic or Technical-Professional;
- Region: North (codes 1, 2, 3, 4, 15), Central (regions 5, 13 [RM]) or South (codes 6, 7, 8, 9, 10, 11, 12, 14); and
- Socio-economic status: Five quintiles of the SES. Quintile A defines lowest SES and Quintile E defines highest SES.

Table 153 shows n-counts of background variables for university-admitted applicants from 2004 through 2011. The table also shows the corresponding percentages (within parenthesis) based on the population of university applicants for each admission year. The n-counts are final reached after screening out incomplete data records and selecting careers with at least 15 valid records. A minimum of 15 records per career was chosen to estimate prediction validity indices with minimal error, retaining as many careers as

possible, and allowing positive degrees of freedom for regression analyses. This rule of thumb has been used internationally in prediction validity studies (Mattern et al., 2008).

Some important patterns can be observed from Table 153. The ratio of Male-Female admitted applicants has remained about the same across admissions years; whereas the ratio of Subsidized-Private and Subsidized-Municipal has shown differences over time . The ratio indicates larger number of students being admitted into universities over the years. For example, for the admission processes 2010 and 2011, there were 23,032 and 22,827 admitted applicants from Subsidized high schools, respectively. For the same admission years, the numbers of admitted applicants from municipal schools were 15,018 and 14,075, respectively, and the numbers of admitted applicants from private schools were 10,489 and 10,541, respectively. Students graduating from the Scientific-Humanistic curricular branch have stronger presence in universities relative to students from the Technical-Professional curricular branch. Over the admission years, about 85% of university-admitted students graduated from the Scientific-Humanistic curricular branch. For example, for the admission processes 2010 and 2011, there were 42,789 and 42,299 admitted students from the Scientific-Humanistic curricular branch, respectively, and 5,750 and 5,144 admitted students from the Technical-Professional curricular branch. Regarding the variable region, central and south regions showed larger percentage of university students in the sample, followed by the north region. The socio-economic status the three upper quintiles (Quintiles C, D, and E) have historically found among university-admitted applicants. Interestingly, for the 2011 admissions year, the number of universities contributing data to MINEDUC databases used in this study was low. Because of the low n-counts and to preserve accuracy of reported statistics, it is recommend exerting caution when interpreting results.

Table 153: N-count (percentage) distribution by university admitted applicants' demographic variables and admission year

	2004	2005	2006	2007	2008	2009	2010	2011
Gender								
Male	16584 (11)	18478 (11)	19638 (11)	21408 (10)	22465 (10)	24710 (10)	24710 (10)	9872 (4)
Female	17405 (11)	18584 (11)	20268 (11)	22088 (10)	22461 (10)	24019 (10)	24019 (10)	9689 (4)
Type of High School								
Private	7453 (5)	8137 (5)	8628 (5)	9039 (4)	9570 (4)	10489 (4)	10541 (4)	5208 (2)
Subsidized	14237 (9)	15764 (9)	17331 (10)	19731 (9)	21003 (10)	23032 (10)	22827 (9)	8824 (4)
Municipal	12167 (8)	13035 (8)	13797 (8)	14571 (7)	14199 (7)	15018 (6)	14075 (6)	5414 (2)
Curricular branch								
Scientific-Humanistic	29079 (19)	31838 (19)	34593 (20)	37313 (18)	39033 (18)	42789 (18)	42299 (17)	17659 (7)
Technical-Professional	4778 (3)	5098 (3)	5163 (3)	6028 (3)	5739 (3)	5750 (2)	5144 (2)	1787 (1)
Region								
North	4186 (3)	4647 (3)	4793 (3)	4948 (2)	5293 (2)	5674 (2)	5838 (2)	1587 (1)
Central	17289 (11)	18585 (11)	20487 (12)	22033 (10)	22909 (11)	24612 (10)	25203 (10)	13377 (5)
South	12476 (8)	13787 (8)	14583 (8)	16472 (8)	16699 (8)	18413 (8)	16607 (7)	4578 (2)
Socioeconomic Status								
Quintile A	4527 (3)	4527 (3)	6062 (3)	7746 (4)	7825 (4)	4900 (2)	5712 (2)	2087 (1)
Quintile B	7452 (5)	7452 (4)	8679 (5)	8977 (4)	10290 (5)	5058 (2)	5261 (2)	1808 (1)
Quintile C	5093 (3)	5093 (3)	5483 (3)	5920 (3)	6646 (3)	6811 (3)	7507 (3)	2921 (1)
Quintile D	3588 (2)	3588 (2)	3653 (2)	3770 (2)	4008 (2)	6436 (3)	7182 (3)	2924 (1)
Quintile E	3558 (2)	3558 (2)	3271 (2)	3431 (2)	3681 (2)	9198 (4)	9967 (4)	4265 (2)

(Note: It is recommended exerting caution when interpreting results for the 2011 admission process. The year is listed in compliance with contract expectation. Percentages are based on the population of university applicants in each admission year.)

Predictor Measures

In Chile admissions decisions are made at the career level within a university with a postulation score, which combines selected PSU scores and *Notas de la Enseñanza Media* (NEM). The admission criteria comprise PSU scale scores and NEM utilizing weights defined by each career within university under policy guidelines defined by the *Consejo de Rectores de Universidades Chilenas* (CRUCH).

The PSU battery is composed of six separate tests: (1) Language and Communication, (2) Mathematics, (3) History and Social Sciences, (4) Biology, (5) Physics, and (6) Chemistry. PSU tests scores are reported on the PSU scale, which has a mean of 500 and a standard deviation of 110 points. The PSU scale score ranges from 150 to 850 points. Higher scores on PSU scale indicate higher test performance. DEMRE extracted data sets with PSU test scores and applicants' survey information.

NEM provide information on high school grade point average of university applicants. NEM are reported in scale scores through a set of norms developed with 2003 admission process. NEM norms were developed for the Scientific-Humanistic (morning and afternoon) and Technical-Professional curricular branches. The NEM scale has a mean of 500 and a standard deviation of 100 points. Higher scores on NEM indicate higher high school academic performance. DEMRE provided applicants NEM along with PSU test scores and applicants' survey information.

Applicants' ranking depicts relative standing of applicants among school peers. This variable is not part of the admission criteria but it was added to the study. Ranking is derived from NEM and reported in a 1 to 100 point scale. Higher score on ranking indicates lower standing. The direction of the scale was reversed to align to the interpretation of PSU and NEM scales. The reversing of the scale does not change rank order of students. It is intended to change direction. On the reversed scale, high ranking indicates high academic standing relative to peers.

MINEDUC provided data sets with applicants' ranking information. To calculate the relative position of a student with respect his or her peers, MINEDUC did the following.

- The ranking considers the High School institution from which the student graduated, regardless of whether the Secondary Education curriculum was wholly or partially implemented in that establishment.
- The relative position of a student compared to all alumni in a given year from the same educational institution.
- The average High School grades for each student was calculated by averaging the overall averages of each year (1st to 4th year of high school or equivalent in adult education).
- For purposes of determining the final average approximations were made to one decimal place, whereas the second decimal, so if this is not more than four, the first decimal is retained, and if it is equal or greater than five, an approximation is made to the number immediately above.
- All High School grades that a student has between 2003 and 2011 were considered, regardless of whether or not the student passed the grade level.
- If a student passed more than once the same grade level, that might be the case of "voluntary repetition," only the mark of the first time the grade was passed is considered.

Finally, it is important to note that in the calculation of those government-issued RUT numbers below one million, more than 100 million (for foreigners without RUT) or had a pattern (11111111-1,

22222222-2, etc.) were not considered. Even though some numbers were in the valid RUT range, they were excluded because they are not attributable to a particular individual. (MINEDUC, 2012)

Criterion Measures

The study involved three university admissions outcomes: (1) first-year university grade point average (FYGPA), (2) second-year university grade point average (SYGPA), and (3) university completion. FYGPA and SYGPA scores range from 1 to 7. Higher scores are indicative of higher university grade point average. University completion was created as a dichotomy of the variable students' university academic standing. The variable has six categories depicting different enrollment status (e.g., regular, suspended, completed course work, completed dissertation). After inspection of distributions of categories and to achieve large enough n-counts, university completion was defined as achieving a standing of university course completion. That is, "university completion" comprises students with completed course work and students who have successfully defended their thesis or any other graduation requirement. A score of one on university completion indicates a standing of at least completion of university courses. A score of zero depicts the other categories. MINEDUC provided data sets collected from universities belonging to CRUCH and eight affiliated private universities.

Statistical Analyses

Statistical modeling relies on complex models to account for a wide array of test performance and past academic performance for differential prediction studies (Arce-Ferrer & Borges, 2007; Roberts & Noble, 2004; Young, 2001). We are using well-regarded methodology which has been used for the SAT (Bridgeman, B., McCamley-Jenkins, L., & Ervin, N., 2000; Burton, N., & Ramist, I., 2001; Kobrin, J., Patterson, B., Shaw, E., Mattern, K., & Barbuti, S., 2008) and ACT (Roberts & Noble, 2004; Maxey & Sawyer, 2001) and for the PSU in Chile (Comité Técnico Asesor, 2010).

For predictive validity, descriptive statistics were summarized for predictor and criterion variables. The prediction validity coefficient, r_{xy}, was computed with the Pearson correlation coefficient to study the association between predictor measures and criterion measures. The analyses were performed by career within university. The uncorrected prediction validity coefficients (r_{xy}) were corrected for restriction of range (Gulliksen, 1987) to render r^*_{xy}. This type of range-restriction correction has been used internationally in prediction validity of university admissions tests (Kobrin et al., 2008). In this model, selection takes place on either predictor (i.e., x-variable) or criterion (i.e., y-variable), the unrestricted variance is known for the selection variable (i.e., variance on x from population of university applicants). The range-restriction formula is shown below:

$$r^*_{xy} = \frac{\left(\frac{S_x}{s_x}\right) r_{xy}}{\sqrt{1 + r^2_{xy}\left(\frac{S^2_x}{s^2_x} - 1\right)}}$$

In the above equation:

x and y depict predictor (e.g., PSU Mathematics score) and criterion (e.g., first year university GPA) variables, respectively;

r_{xy}^{*} stands for corrected prediction validity;

r_{xy} indicates uncorrected prediction validity coefficient;

S_x^2, S_x define variance and standard deviation, respectively, of predictor variable for the unrestricted population (i.e., population of students seeking university admission);

s_x^2, s_x define variance and standard deviation, respectively, of predictor variable for the restricted population (i.e., population of students admitted into university studies).

Table 154 shows unrestricted standard deviation and variance for PSU Mathematics scores across administration years. (Note: Appendix O shows the same information for all predictor measures.)

Table 154: PSU Mathematics Score Unrestricted Standard Deviation and Variance by Admission Year

Year	N	Mean	Standard Deviation	Variance
2004	153383	499.99	109.46	11981.25
2005	169376	500.55	110.14	12129.81
2006	176314	500.61	110.24	12153.68
2007	211261	500.31	109.55	12000.53
2008	216892	500.36	109.87	12072.09
2009	242130	500.20	109.59	12009.52
2010	251634	500.79	110.77	12270.17
2011	250758	501.07	111.27	12380.78
2012	231140	500.36	109.73	12040.97

After correcting for range restriction, prediction validity indices were averaged utilizing Hunter and Schmidt's approach (1990). The approach consists in weighting the corrected validity coefficients with their sample size within careers to take into account random sampling error. The process weight prediction validity coefficients when summarizing outcomes from prediction validity studies is used internationally on university admissions research (Burton & Ramist, 2001). The formula is shown below:

$$WeightedAv\ erage\left(r_{xy}^{*}\right)=\frac{\sum_{i} n_i r_{xy}^{*}}{\sum_{i} n_i}$$

Where *i* indexes career, r_{xy}^{*} stands for corrected prediction validity index for a given career, *n* represents sample size, and ∑ is a summation operator over careers.

To study prediction validity, simple regression analyses were performed within career with at least 15 valid records. For these analyses the following regression models were adjusted to the data set for each admission year:

Model 1: Criterion: First year university grade point average
Predictors: PSU test scores and high school grade point average (NEM), and ranking
Model: Single linear regressions

Model 2: Criterion: Cumulated second year university grade point average
Predictors: PSU test scores, high school grade point average (NEM), and ranking
Model: Single linear regressions

Model 3: Criterion: University graduation
Predictors: PSU test scores, high school grade point average (NEM), and ranking
Model: Single linear logistic regressions

To document predictive performance of the ranking variable, multiple linear regressions were performed within career with at least 15 valid records. The intention of these analyses was to obtain contribution of the ranking variable on reduction of university outcome variance after controlling for the PSU tests scores and NEM. For these analyses models 1-3 were revised to add the variable ranking as shown below to devised models 1R to 3R. The revised models were adjusted to the career data set and summarized over admission years utilizing Hunter and Schmidt's approach.

Model 1R: Criterion: First year university grade point average
Predictors: PSU test scores and high school grade point average (NEM). The model was adjusted twice with and without the variable ranking
Model: Multiple linear regressions

Model 2R: Criterion: Cumulated second year university grade point average
Predictors: PSU test scores and high school grade point average (NEM). The model was adjusted twice with and without the variable ranking
Model: Multiple linear regressions

Model 3R: Criterion: University graduation
Predictors: PSU test scores and high school grade point average (NEM). The model was adjusted twice with and without the variable ranking
Model: Multiple linear logistic regressions

PSU differential predictive validity analyses were performed with a methodology that involves career as unit of analysis, single and multiple-predictors linear regression models, and analyses standardized prediction error disaggregated by subpopulation of interest (See Figure 50).

In a second tier of analyses, estimates of standardized residuals were computed within careers and disaggregated by subpopulations. For summative purposes individual students'

standardized residuals were averaged across careers and admission years before disaggregating them by subpopulations. Negative residuals are indicative of over-prediction of actual first-year cumulative university grade point average, and positive residuals indicate under-prediction of actual first-year cumulative university grade point average. Qualitative interpretations of residuals will be made with reference to the subgroups.

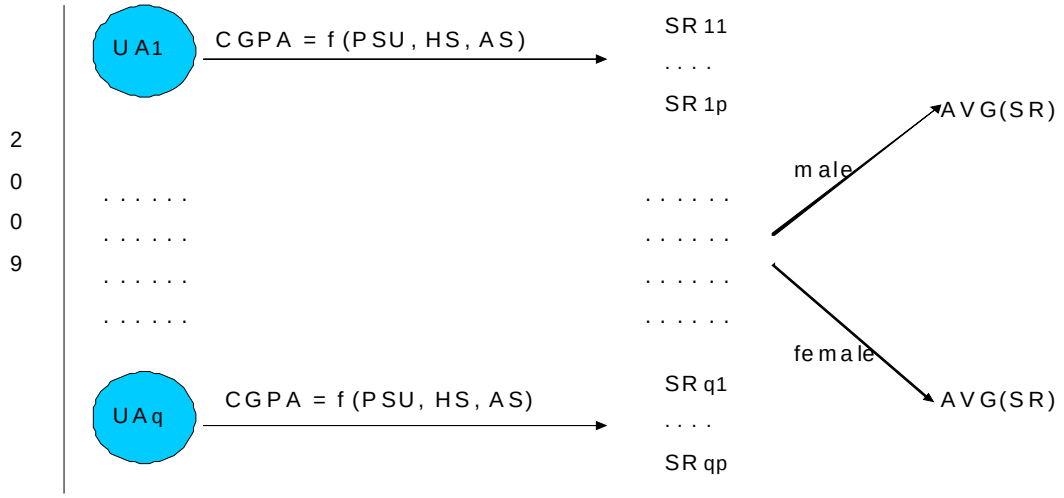


Figure 50: Differential prediction analyses for male-female subpopulations and year 2009

In Figure 50, UA stands for unit of analysis which is the career, *q* indexes number of units of analyses, CGPA defines cumulative university grade point average which can be FYGPA, SYGPAS or University completion, *f* stands for single or multiple linear regression function, PSU represents PSU test scores, HS stands for high school grade point average which is NEM, AS stands for ranking, SR defines standardized residuals, *p* indexes number of students within a unit of analysis, and AVG indicates average operator from Hunter and Schmidt.

The specific steps performed for documenting differential prediction validity are listed below:

STEP 1: Select unit of analyses (careers) with the following characteristics:

- Do for admission process 2004 (Repeat for other years of admission)
- Select records valid records showing data for all predictors and the criterion measures
- Check each career has at least 15 valid records after subsetting by the admissions process and valid predictor and criterion

STEP 2: Within each career, adjust regression models 1 to 3 and output standardized residuals

STEP 3: Within each career, average standardized residuals by subgroup of interest (i.e. gender, region, etc.)

STEP 4: After mean standardized residual by subgroup is calculated, weighted averages across schools were generated using the Hunter and Schmidt's procedure mentioned above.

(Note: Predictive and incremental validity and differential prediction bias analyses were also performed by career type. For each of the careers defined by DEMRE, the models above described were adjusted to the data sets and summarized following the procedures already described in the above sections.)

RESULTS

This section summarizes in four sections the data analysis results performed with cohorts of university admitted applicants. The first section provides a descriptive summary of the predictor measures and the criterion measures. The second section brings a summary of the predictive validity findings for FYGPA, SYGPA, and university completion. The third section summarizes findings on the incremental predictive validity of the variable ranking on FYGAP, SYGPA, and university completion. The final section is reserved to summarize prediction bias results for the relevant demographic variables.

Descriptive Statistics

Table 155 shows descriptive statistics of the predictor measures for 2004 through 2011 admission process for the population of university admitted applicants. The information shown in the table arises from final n-counts after screening out incomplete data records and selecting careers with at least 15 records. For the 2004 admission process, a total of 27 universities, 735 careers, and approximately 40,000 admitted applicants contributed PSU Language scores. The average PSU scale score for Language was 604.83 points (SD=93.07). The average scale score has shown a positive increase across years. For example, in Mathematics, the average scale score ranged from 594.67 scale score points (in 2004) to 619.61 scale score points (in 2011). It is also relevant to highlight the restriction on the standard deviation of the predictor measures in the group of admitted applicants. As expected, there is less variability in PSU test scores in the population of admitted applicants relatively to the population of applicants. The table shows restricted standard deviation of PSU Mathematics scale score in 2007 to be 79 points. From Table 219, the corresponding unrestricted standard deviation was 109.55 scale score points. The difference between restricted and unrestricted standard deviations is expected to be due to the selection process that took place on the group of admitted university students.

Table 155: Descriptive Statistics of Admitted Applicants' Predictor Measures by Admission Year

	2004	2005	2006	2007	2008	2009	2010	2011*
PSU Language								
N Universities	27	27	27	27	27	27	25	13
N Careers	735	800	844	895	900	932	846	320
N Students	33975	37606	41125	45206	47543	51045	50084	19972
Mean	604.83	592.89	591.8	598.85	597.11	602.85	606.87	620.29
Std. Dev.	93.07	81.99	81.8	79.93	80.4	79.44	77.91	76.12
PSU Mathematics								
N Universities	27	27	27	27	27	27	25	13
N Careers	735	800	844	895	900	932	846	320
N Students	33975	37606	41125	45206	47543	51045	50084	19972
Mean	594.67	597.73	596.41	602.83	602.23	608.53	610.86	619.61
Std. Dev.	84.69	82.01	82.96	79	80.54	78.21	81.12	81.73
PSU History								
N Universities	27	27	27	27	27	27	25	13
N Careers	475	553	590	623	633	676	640	247
N Students	19256	21578	23497	25297	26131	27652	27599	11819
Mean	596.43	589.90	589.02	597.67	595.92	601.63	606.30	617.72
Std. Dev.	97.70	90.90	90.61	87.71	88.17	87.36	88.14	86.05
PSU Science								
N Universities	26	26	27	27	27	27	25	13
N Careers	529	569	624	666	677	712	656	221
N Students	21434	23623	26341	29066	31272	34518	33891	12166
Mean	566.52	577.68	577.23	581.01	582.44	587.97	589.21	600.92
Std. Dev.	86.39	90.34	90.77	88.29	87.44	87.12	86.56	89.07
NEM								
N Universities	27	27	27	27	27	27	25	13
N Careers	734	796	841	890	894	925	837	317
N Students	33917	37209	40267	44151	46026	49685	48704	19768
Mean	627.68	628.52	623.63	622.13	614.78	614.63	608.81	608.08
Std. Dev.	88.88	89.52	89.81	90.19	91.74	91.59	93.13	92.25
Ranking								
N Universities	27	27	27	27	27	27	25	13
N Careers	523	743	816	877	882	918	838	313
N Students	17509	30136	36352	41872	44863	48644	47849	19140
Mean	33.12	34.01	34.18	33.97	35.48	35.31	36.62	37.32
Std. Dev.	25.82	26.36	26.32	26.41	26.76	26.72	27.19	27.04

(* Note: For the 2011 admission process, the number of universities contributing data to MINEDUC database was about half of the number observed for other years. It is recommended exerting caution when interpreting results involving the 2011 admission process. The year is listed in compliance with contract expectation.)

Table 156 shows descriptive statistics of criterion measures for university students admitted in 2004 through 2011 processes. The table shows sample sizes after screening out incomplete data records, selecting careers with at least 15 records, and merging data with PSU scores. For the 2004 admission process, a total of 27 universities, 672 careers, and approximately 29,000 university students contributed valid FYGPA. The average FYGPA was 4.72 (SD=0.92). The FYGPA appears to be stable across years. For SYGPA the numbers of universities and careers contributing data to the study are robust for admission processes 2004 through 2010. For university completion analyses admission process 2004 through 2006 contributed the data for the analyses, as expected from lag in time between university admission and university completion.

Table 156: Descriptive Statistics of Admitted Applicants' Criterion Measures by Admission Year

	2004	2005	2006	2007	2008	2009	2010	2011*
University FYGPA (1-7 scale)								
N Universities	27	25	25	25	25	25	23	2
N Careers	672	709	773	825	829	869	784	9
N Students	29201	31326	35872	39645	41779	45242	43017	584
Mean	4.72	4.69	4.72	4.70	4.66	4.66	4.62	4.69
Std. Dev.	0.94	0.94	0.93	0.95	0.94	0.95	0.97	0.76
University SYGPA (1-7 scale)								
N Universities	27	26	25	25	25	25	2	1
N Careers	647	707	747	784	791	804	46	5
N Students	26378	29237	32322	35377	37080	38147	2737	154
Mean	4.80	4.79	4.81	4.79	4.75	4.77	5.02	5.16
Std. Dev.	0.87	0.87	0.86	0.87	0.86	0.86	0.92	0.37
University Completion (1=yes, 0= no)								
N Universities	27	27	27	N/A	N/A	N/A	N/A	N/A
N Careers	742	805	856	N/A	N/A	N/A	N/A	N/A
N Students	34846	38307	41777	N/A	N/A	N/A	N/A	N/A
Mean	0.45	0.38	0.26	N/A	N/A	N/A	N/A	N/A
Std. Dev.	0.50	0.49	0.44	N/A	N/A	N/A	N/A	N/A

(Note: For the 2011 admission process there were two universities with reported FYGPA in MINEDUC database. For 2011 admission process there was one university with reported SYGPA in MINEDUC database. For 2010 admission process and SYGPA there was two universities. For university completion, data waves from admission process spanning 2004 through 2006 were used. N/A stands for not applicable.)

Predictive Validity

The main statistical approach used in this study was the linear regression model to study strength of association between predictor measures (PSU test scores, NEM, ranking) and criterion measures (university FYGPA, SYGPA, and completion). A correlation of 1.0, for instance, implies a perfect linear relationship between predictor measures and a criterion measure. On the other hand, a correlation of 0.0 indicates a lack of linear relationship. Negative magnitudes of correlation coefficients may also plausibly be experienced; however, they are rarely found in prediction validity studies with large sample sizes. Gauging the size of predictive validity indexes can be done utilizing rules of thumb available from literature (Cohen 1988). Absolute value of correlations of approximately 0.10 are considered small, those of approximately 0.30 are considered medium, and those of approximately 0.5 and above are considered large. Predictive validity indexes can be also evaluated relatively to benchmarks derived from international testing programs. Internationally, predictive validity of university admissions tests with first-year university grades ranged from 0.45 to 0.55 (Beatty, Greenwood & Linn, 1999; Kobrin et al., 2008; Young, 2001).

Table 157 shows average prediction validity indices for PSU test scores, NEM, and Ranking on university FYGPA over admission years. The table reports validity coefficients corrected by range restriction and summarized with Hunter and Schmidt's approach described in the methodology section. Regarding predictive validity of single PSU predictors with university FYGPA, correlations corrected for range restriction, ranged from 0.11 (PSU History and Social Sciences for 2006 admission) to 0.40 (PSU Mathematics for 2009 admission and PSU Science for 2005 admission process). The median predictive validity across PSU tests over 2004-2010 admission period was 0.34 (Note: information from the 2011 data set was dropped due to aberrant results.) When analyzing individual PSU test prediction validity over years, it becomes evident of the low predictive power of PSU Language and Communication and History and Social Sciences tests. Systematically prediction validity index showed magnitudes indicative of small relationship with university FYGPA. On the other hand, PSU Mathematics and Science tests showed medium values of predictive validity. In none instance PSU tests achieve prediction validity indexes closer to the lower bound observed internationally. Results by type of career showed similar prediction patterns of FYGPA to those from the overall analyses. For example, PSU Mathematics and Science scores and high school academic performance (NEM and ranking) showed larger predictive capacity than PSU Language and Communication and History and Social Sciences scores. Predictive validity tables for university outcomes by each career type are available in Appendices P, Q, and R, which show univariate predictive validity results for FYGPA, SYGPA and university academic outcomes, respectively.[16] The analyses by type of career showed smaller size of predictive validity coefficients for PSU test scores.

[16] The PSU test battery is composed of four standardized instruments: two mandatory tests ("Mathematics" and "Language and Communication") and two optional tests ("Science" and "History and Social Sciences"). By policy, one of the two optional tests must be included as a selection factor for each university career. The student n-counts in the table(s) show the number of applicants on the mandatory assessments across careers and universities for a particular year and career family. Therefore, the number of applicants taking either one or both of the optional tests is less than or equal to the number of those reported in the tables. N.B., the data sets do not distinguish by career type the optional test that was used for admissions in each particular career group in each particular university.

Regarding predictive validity of single high school predictors (NEM and ranking) with university FYGPA, predictive validity coefficients, corrected for range restriction, ranged from 0.27 (rank for 2010 admission) to 0.37 (NEM for 2005 admission). Two aspects are important to highlight from the findings. One is the strength of relationship of high school predictors. High school performance (NEM and Ranking) showed comparable prediction performance with best predictive PSU test scores (Mathematics and Science). Internationally similar findings have been reporting for university admissions tests (Kobrin et al., 2008; Linn 1990). The other pattern is the lower predictive validity magnitude relatively to international standards. Internationally high school records have been found largely correlated with university FYGPA. The median correlation among hundreds of studies is about 0.48 (for SAT) and 0.50 (for ACT) (Linn, 1990). As expected from their common origin, the ranking variable exhibited a performance similar to the NEM. Results by type of career showed patterns of predictive validity of NEM and ranking that were similar to those from the overall analyses. For example, NEM and ranking showed larger predictive validity coefficients than PSU test scores for most types of careers.

Table 157: Average Pearson correlations (corrected by range restrictions) between Predictor Measures and University FYGPA by Admission Year

Sample Sizes				PSU Tests				High School	
Year	University	Career	Student	Language and Communication	Mathematics	History and Social Sciences	Science	NEM	Rank
2004	27	662	28400	0.13	0.38	0.14	0.37	0.36	0.36
2005	25	705	30864	0.18	0.35	0.12	0.40	0.37	0.33
2006	25	766	35427	0.19	0.35	0.11	0.35	0.34	0.30
2007	25	819	39264	0.18	0.36	0.12	0.33	0.35	0.28
2008	25	827	41488	0.18	0.34	0.13	0.38	0.35	0.28
2009	25	864	44978	0.18	0.40	0.12	0.35	0.35	0.28
2010	23	781	42824	0.18	0.34	0.12	0.36	0.35	0.27
2011	2	9	584	0.04	-0.05	-0.04	0.34	0.40	N/A

(Note: The complete tables and unrestricted standard deviations used to correct for range restriction are located in Appendix O. Year 2011 is shown in compliance with contract expectation. Caution is recommended making inference with that small number of universities. N/A stands for not available.)

Table 158 shows average prediction validity indices for PSU test scores, NEM, and Ranking on university SYGPA over admission years. The table reports validity coefficients corrected by range restriction and summarized with Hunter and Schmidt's approach described in the methodology section. Regarding PSU predictors, predictive validity ranged from 0.08 (PSU History for 2007 admission) to 0.39 (NEM for 2004 and 2005 admission years). The median predictive validity was 0.29 across PSU test and over admission years. Relative to the FYGPA findings, SYGPA predictive validity indices went downward. When analyzing individual PSU test prediction validity over years, it becomes evident the diminution in power of PSU tests to predict university SYGPA. PSU Language and History tests showed the lowest predictive prediction validity coefficients. Systematically, these two tests experienced prediction validity indices indicative of small relationship with university SYGPA. On the other hand, PSU Mathematics and Science test showed borderline medium values of predictive validity. In none instance PSU tests achieve prediction validity indexes closer to the lower bound observed internationally. Results by type of career showed similar prediction patterns of SYGPA to those from the overall analyses. For example, PSU Mathematics and Science scores and high school academic performance (NEM and ranking) showed larger predictive capacity than PSU Language and Communication and History and Social Sciences scores. Predictive validity tables for university outcomes by each career

type are available from Appendix R. The analyses by type of career showed smaller size of predictive validity coefficients for PSU test scores.

Regarding predictive validity of single high school predictors (NEM and ranking) with university SYGPA, predictive validity coefficients, corrected for range restriction, ranged from 0.24 (rank for 2010 admission) to 0.39 (NEM for 2004 admission). Two aspects are important to highlight from the findings. One is the retention of prediction power of high school predictors over the lost of prediction power of PSU test scores. High school performance (NEM and Ranking) showed higher prediction performance with best predictive PSU test scores (Mathematics and Science). Performance in high school captures multiple variables such as students' motivation and persistence over a larger period of time than taking PSU university admissions tests with multiple choice items tapping content measurable with paper and pencil tests. As expected from their common origin, the raking variable exhibited similar performance to NEM. Results by type of career showed patterns of predictive validity of NEM and ranking that were similar to those from the overall analyses. For example, NEM and ranking showed larger predictive validity coefficients than PSU test scores for most types of careers.

Table 158: Average Person correlations (corrected by range restrictions) between Predictor Measures and University SYGPA by Admission Year

Sample Sizes				PSU Tests				High School	
Year	University	Career	Student	Language and Communication	Mathematics	History and Social Sciences	Science	NEM	Rank
2004	27	641	25736	0.12	0.28	0.09	0.29	0.39	0.37
2005	26	697	28642	0.17	0.27	0.09	0.33	0.39	0.35
2006	25	740	31946	0.18	0.28	0.10	0.28	0.38	0.33
2007	25	779	35088	0.17	0.28	0.08	0.25	0.37	0.29
2008	25	786	36811	0.15	0.26	0.10	0.28	0.36	0.30
2009	25	800	37939	0.15	0.31	0.11	0.28	0.36	0.30
2010	2	46	2737	0.15	0.12	0.11	0.22	0.33	0.24

(Note: The complete tables and unrestricted standard deviations used to correct for range restriction are located in Appendix O. Year 2010 is shown in compliance with contract expectation. Caution is recommended making inference with that small number of universities.)

The ability to forecast long-term university success is an important piece of evidence supporting prediction validity of standardized admission tests. There is a relatively slow pace in collecting long term information of test scores predictive validity. Forecasting long term success in university requires elapsed time periods from freshman years to graduation before data becomes accessible to researchers. Nonetheless, there is a small number of studies on long term university success in which university admissions tests were analyzed (Burton & Ramist, 2001) and from which several lessons were learned. Overall analyses of SAT and long term success have found to have been instrumental in understanding the origin for the small declines in predictive validity of tests scores and high school performance. Validity coefficients tend to be lower when university courses cover a wide range of content from the academic to the practical and grading standards differ among university courses. Results by type of career showed similar prediction patterns of University Completion rates to those from the overall analyses. For example, PSU Mathematics and Science scores and high school academic performance (NEM and ranking) showed larger predictive capacity than PSU Language and Communication and History and Social Sciences scores. Predictive validity tables for university outcomes by each career

type are available in Appendix R. The analyses by type of career showed smaller size of predictive validity coefficients for PSU test scores.

Table 159 shows average prediction validity indices for PSU test scores, NEM, and Ranking on university completion over admission years. The table reports validity coefficients corrected by range restriction and summarized with Hunter and Schmidt's approach described in the methodology section. Regarding PSU predictors, predictive validity ranged from 0 (PSU History for 2006 admission year) to 0.09 (PSU Science for 2005 admission year, and PSU Mathematics for 2004 admission year). The median predictive validity was 0.03 across PSU tests and over admission years. For individual PSU tests prediction validity indices for PSU Language and Communication and History and Social Sciences slide to magnitudes near null associations. PSU Mathematics and Science tests showed small magnitudes in their prediction validity coefficients. The median predictive value for the two tests over admission years was 0.08. The lower correlations are expected since university completion can be influenced by nonacademic factors such as finance, motivation, personal decisions, and social adjustment. Nevertheless, PSU tests capability for predicting long term university success slide below the lower bounds reported from international literature on SAT test (about 0.40) (Burton & Ramist, 2001). Results by type of career showed patterns of predictive validity of NEM and ranking that were similar to those from the overall analyses. For example, NEM and ranking showed larger predictive validity coefficients than PSU test scores for most types of careers.

Regarding predictive validity of single high school predictors (NEM and ranking) with university completion, predictive validity coefficients, corrected for range restriction, ranged from 0.04 (rank for 2006 admission year) to 0.11 (NEM for 2004 admission year). High school performance (NEM) showed higher prediction performance relatively to performance of the best predictive PSU test scores (Mathematics and Science). The ranking variable slide down a bit; though it remained with sizes comparable to NEM. University persistence can be influenced by nonacademic factors; henceforth low correlations can be expected when predicting long term university success with academic predictors.

Table 159: Average Pearson correlations (corrected by range restrictions) between Predictor Measures and University Completion

Sample Sizes				PSU Tests				High School	
Year	University	Career	Student	Language and Communication	Mathematics	History and Social Sciences	Science	NEM	Rank
2004	27	735	33975	0.02	0.09	0.03	0.08	0.11	0.09
2005	27	800	37606	0.03	0.08	0.02	0.09	0.10	0.08
2006	27	844	41125	0.02	0.03	0.00	0.03	0.06	0.04

(Note: The complete tables and unrestricted standard deviations used to correct for range restriction are located in Appendix O.)

Incremental Predictive Validity For Ranking

Multiple linear regression was the main statistical approach for investigating incremental predictive validity for the variable ranking. The base model with PSU test scores and NEM was fitted to the criterion measures (university FYGPA, SYGPA, and completion) to estimate squared multiple correlation coefficients. Following, the ranking variable was added to the base model to estimate squared multiple correlation coefficients with the revised model.

The differences on squared multiple correlations between base and revised models were used to index amounts of increase in prediction brought by ranking.

Table 160, Table 161 and Table 162 show incremental predictive validity of ranking based on multiple squared correlations for base and revised models. The multiple regression coefficients were not corrected by range restriction because weighted covariance matrices needed to perform the corrections cannot be estimated for the unrestricted population (Dunbar & Linn, 1991; *Comité Técnico Asesor*, 2010). The results shown in the tables are valuable to explore reduction of uncertainty on criterion measure due to the ranking variable.

After controlling for PSU test scores and NEM, ranking contributed in reducing uncertainty on FYGPA and SYGPA. The amount of variance reduction of university FYGPA and SYGPA was of no more than 4%. Ranking variance reduction ranged from 2% to 4% with median of 3%. Interestingly, the variable ranking contributed about 6% of reduction of uncertainty for long term university success (university completion) over and beyond PSU test scores and NEM. Internationally, in the context of SAT, uncorrected correlation of 0.36 was reported between high school record (self reported) and university FYGPA. This correlation accounted for by 13% variance reduction of FYGPA (Kobrin, et al., 2008).

The ranking variable accounted for by 6% to 7% of variance of University Completion after controlling by other predictor measures in the model. The median variance reduction of ranking over admission years was 6%. Recent studies predicting university completion with SAT reported uncorrected correlations between high school record (grade point average or class ranking) a university completion of 0.29 (Burton & Ramist, 2001). The correlation translates into an about 8.4% in variance reduction of university completion. These correlations suggest that there is an academic component to university completion arising from university admission and highs school grade point average. The median R-squared for university completion over years for the base model was 0.15 and for the revised model was 0.21. The correlation with university completion for PSU tests and NEM (base model) was larger than the correlation of either NEM or Ranking alone. The incremental validity analyses per type of career are available in Appendix U. At the career level, the incremental predictive validity of ranking was smaller than from the general analyses.

Table 160: Average R-square for Base and Revised Models and FYGPA

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	26	445	14305	0.21	0.25	0.04
2005	25	638	24422	0.20	0.23	0.03
2006	25	732	30374	0.18	0.21	0.03
2007	25	785	34904	0.17	0.20	0.03
2008	25	794	36994	0.16	0.19	0.03
2009	25	836	40885	0.17	0.19	0.02
2010	23	754	2149	0.16	0.18	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 161: Average R-square for Base and Revised Models and SYGPA

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	26	411	12716	0.20	0.25	0.04
2005	26	633	22746	0.19	0.22	0.04
2006	25	692	27294	0.18	0.21	0.03
2007	25	757	31440	0.16	0.19	0.03
2008	25	749	32795	0.15	0.18	0.03
2009	25	774	34617	0.16	0.18	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 162: Average R-square (Cox-Snell) for Base and Revised Models and University Completion

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	27	390	13138	0.17	0.24	0.07
2005	27	583	22818	0.15	0.21	0.06
2006	27	577	24371	0.14	0.20	0.06

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Differential Predictive Validity

Predictive validity studies have focused on understating adequacy of prediction models across subgroups of test takers. The term differential predictive validity is used to describe degrees of prediction model generalizations across subpopulations of test takers. In the ideal world university admissions tests are develop to forecast future academic success in university in such a way that actual university grades and model predicted grades are the same. That is, when prediction models are derived from a group of males and females, applicants' gender would make no difference on predicted future performance. That is, a common prediction model would suffice to explain future academic performance for male and female applicants. Under these circumstances, predicted and actual university grades would be the same for male and female subgroups. In real world, however, a prediction equation may hold differently for subpopulations and yield expected performance that would underpredict or overpredict actual academic performance of subpopulations.

Internationally, Ramist et al., (1994) found that a base prediction model of university FYGPA with SAT scores and high school GPA underpredicted female university grades by 0.06. The result indicated that females' predicted university FYGPA were 0.06 grade points lower than their actual grades, based on a 4-point grading scale. For example, a prediction equation would forecast a 3.2 university FYGPA for a group of females when they would average 3.26. More recently, Mattern et al., (2008) reported presence of small differential prediction bias for SAT test scores. The results for gender revealed that SAT tends to underpredict FYGPA for females with mean standardized residuals ranging from 0.07 to 0.11.

Table 163,Table 164 and Table 165 show the average differential prediction validity for PSU tests, NEM, and Ranking and university FYGPA and SYGPA whilst almost null average differential prediction for university completion. The tables report standardized residuals summarized with Hunter and Schmidt's approach described in the methodology section. Tables 11 and 13 showed that PSU tests and high school academic performance under predict short term university outcomes for females and high school graduates from the Technical-Professional curricular branch. For females the magnitude of underprediction ranged from 0.02 (NEM) to 0.10 (PSU History) with a median of 0.07 for FYGPA and SYGPA. For high school graduates from Technical-Professional buildings underprediction ranged from 0 (NEM) to 0.11 (PSU Mathematics) with a median of 0.05 for FYGPA, and from 0.03 (NEM) to 0.12 (PSU Mathematics) for SYGPA, with a median of 0.07.

Differential prediction validity for long term university success was almost negligible for females and graduates from Technical-Professional curricular branch. Whereas females' average under prediction ranged from 0.00 (NEM) to 0.02 (PSU Mathematics, Language and Communication, and History and Social Sciences tests); Technical-Professional high school graduates' average under prediction was 0.02.

Although there may be some particular instances for which underprediction rates observed in this study could change odds of passing a course (e.g., students closer to passing threshold), observed rates have small practical significance for short term university success. For example, a group of females expected receiving a 4.7 university FYGPA would average 4.8, on a 7-points scale.

Prediction bias by type of career showed similar patterns to those observed from the overall analyses. For example, PSU test scores and high school academic performance (NEM and ranking) under predicted university outcomes for female university students. Interestingly, size of prediction bias was larger for several careers than for the overall analyses. Prediction bias tables per type of career are available in Appendix X.

Table 163,Table 164 and Table 165 show smaller differential prediction for the remaining subpopulations on short and long term university success.

Table 163: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Lang. and Comm.	History and Soc. Sciences	Science	NEM	Ranking
Gender:								
Male	1013	128997	-0.09	-0.07	-0.11	-0.07	-0.02	-0.03
Female	1013	127328	0.09	0.07	0.10	0.08	0.02	0.03
SES:								
QA	970	31564	0.04	0.03	0.03	0.03	-0.02	-0.03
QB	970	38748	0.01	-0.01	-0.01	-0.01	-0.05	-0.07
QC	970	32298	-0.01	-0.02	-0.03	-0.01	-0.02	-0.04
QD	970	25088	-0.02	-0.02	-0.03	-0.01	0.00	-0.02
QE	970	29610	-0.01	0.00	0.01	0.00	0.04	0.04
Curricular Branch:								
Scientific-Humanistic	1010	222975	-0.01	0.00	-0.01	-0.01	0.00	0.01
Technical-Professional	1010	32235	0.11	0.04	0.05	0.07	0.00	-0.09
High School Type:								
Municipal	972	82137	0.00	-0.01	-0.02	-0.01	-0.03	-0.07
Subsidized	972	113326	0.01	0.00	0.01	0.01	0.00	-0.02
Private	972	59257	-0.01	0.02	0.02	0.00	0.05	0.12
Region:								
Center	806	134333	0.01	0.01	0.01	0.01	0.03	0.02
North	806	19017	-0.03	-0.03	-0.03	-0.01	-0.06	-0.04
South	806	78682	-0.02	-0.02	-0.02	-0.01	-0.03	-0.02

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within universities.)

Table 164: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Lang. and Comm.	History and Soc. Sciences	Science	NEM	Ranking
Gender:								
Male	979	96725	-0.09	-0.08	-0.11	-0.07	-0.02	-0.04
Female	979	98711	0.09	0.08	0.10	0.08	0.02	0.04
SES:								
QA	940	24245	0.06	0.05	0.05	0.05	0.00	-0.02
QB	940	31203	-0.01	-0.03	-0.02	-0.02	-0.06	-0.07
QC	940	23328	-0.02	-0.02	-0.01	-0.01	-0.02	-0.04
QD	940	16966	-0.02	-0.02	-0.01	-0.01	0.00	-0.01
QE	940	18941	-0.01	0.00	0.00	0.00	0.03	0.05
Curricular Branch:								
Scientific-Humanistic	966	166565	-0.02	-0.01	-0.01	-0.01	0.00	0.01
Technical-Professional	966	25226	0.12	0.07	0.07	0.09	0.03	-0.06
High School Type:								
Municipal	942	63331	0.00	-0.02	-0.03	-0.02	-0.03	-0.07
Subsidized	942	84874	0.00	0.00	0.01	0.01	-0.01	-0.02
Private	942	45505	0.01	0.03	0.04	0.02	0.06	0.13
Region:								
Central	730	99718	0.02	0.01	0.02	0.01	0.03	0.02
North	730	13800	-0.03	-0.03	-0.04	-0.02	-0.07	-0.05
South	730	55002	-0.02	-0.02	-0.02	-0.01	-0.03	-0.02

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within colleges.)

Table 165: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Lang and Comm.	History and Soc. Sciences	Science	NEM	Ranking
Gender:								
Male	920	55021	-0.02	-0.02	-0.02	-0.01	0.00	-0.01
Female	920	56048	0.02	0.02	0.02	0.01	0.00	0.01
SES:								
QA	877	17868	-0.02	-0.02	-0.02	-0.02	-0.03	-0.04
QB	877	24648	0.00	-0.01	-0.01	-0.01	-0.01	-0.02
QC	877	16425	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
QD	877	11066	-0.01	-0.01	0.00	-0.01	-0.01	-0.01
QE	877	10374	0.04	0.04	0.04	0.04	0.05	0.05
Curricular Branch:								
Scientific-Humanistic	900	92160	0.00	0.00	0.00	0.00	0.00	0.00
Technical-Professional	900	15534	0.02	0.01	0.01	0.01	0.01	-0.02
High School Type:								
Municipal	871	24427	-0.01	-0.01	-0.01	0.00	-0.01	-0.02
Subsidized	871	46608	0.00	0.00	0.00	-0.01	-0.01	-0.01
Private	871	37622	0.02	0.02	0.03	0.02	0.03	0.05
Region:								
Central	625	50502	0.01	0.00	0.00	0.00	0.01	0.00
North	625	12658	-0.01	-0.01	-0.01	-0.01	-0.01	0.00
South	625	23682	0.00	0.00	0.00	0.00	-0.01	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within colleges.)

EVALUATION

The purpose of this study was three-fold: (1) document the ability of PSU test scores and academic performance in high school (high school GPA and high school ranking) to predict university students' academic outcomes; (2) document incremental prediction value of the variable ranking; and (3) examine the extent to which PSU test scores and high school academic performance exhibits differential prediction across relevant demographic variables (gender, SES, region, high school curricular branch, type of school). The study relied on three outcome variables: (1) university first-year grade point average, (2) university second-year grade point average, and (3) university graduation[17]. The study relied on longitudinal data sets spanning 2004-2012 university admissions processes. Universities within the CRUCH and the eight affiliated universities contributed data for the cohorts; though data was available for a sizable fraction of universities, but not for all of them. DEMRE provided databases for PSU test scores and high school grade point average (NEM), and MINEDUC provided databases with university outcomes and ranking variable. Linear and logistic regression analyses were run separately for each career within a university and summarized across careers and colleges. Corrections for restriction of range were applied to correlation coefficients to deal with selection of university students.

Results from this piece of research indicated that PSU tests have a degree of prediction of university outcomes, particularly first- and second-year grade point average. The prediction validity indices found, however, were smaller than those reported internationally (Matter, et al., 2008). Additionally, sizes of PSU prediction validity coefficient decreased by career type. The variable ranking contributed to the reduction of uncertainty of university outcomes after controlling for PSU test scores and NEM. The largest amount of variance reduction (7%) happened for university completion. When analyses were performed by career, predictive validity of NEM and ranking variables were little affected relatively to PSU test scores predictive validity coefficients. Finally, prediction bias findings showed under-prediction patterns for females similar to those reported internationally (Matter, et al., 2008). Interestingly the magnitude of prediction bias was smaller than those reported internationally. Graduates from Technical-Professional curricular branch showed under prediction for short term university success outcomes; although the size was, for all practical purposes, small. Analyses by career showed similar patterns of predictive bias to those from the general analyses. Other relevant demographic variables considered in the study resulted in negligible differential prediction bias. In sum, the PSU test scores and high school performance measures appear to result in comparable amounts of differential prediction validity for major demographic variables. However, the magnitude of bias can be larger for some careers.

University admissions programs most often rely on admission criteria involving admission test scores and high school academic performance (either grade point average or class rank). In most occasions admission test scores are given the greatest weight in the selection process. For example, in the United States performance on the ACT and SAT is used to envision how well an applicant is likely to perform in university studies. Validity evidence to support that claim is often supported by predictive validity studies involving admission scores and first year university grade point average. Correlations can vary from 0.45 to 0.55.

There are several reasons for correlations to take different values and research has suggested factors that may be at work to make correlations lower than they actually are.

[17] Predictive and incremental validity and differential prediction bias analyses were also performed by career type.

Practitioners are well aware on the effects of restriction of range in lowering the size of predictive validity correlations. Student admission restricts the range of admission test scores and thus lowers predictive validity because it has only scores and university grades from the pool of admitted students that are used to compute predictive validity coefficients. In addition, university grades are not based on a uniform standard, and they may vary from class to class, from career to career and from university to university. The effect of different grading practices among universities can lower the size of predictive validity indices. This point also raises the importance of reliability of first year university grade point average. Low reliability of criterion measures (e.g., first-year grade point average) would attenuate correlations between predictor measures and criterion. More research is recommended along this line to understand university grading practices and the precision of grades to ascertain the degree of impact on prediction validity of PSU test scores. The same recommendation for future research applies to high school grading practices and the precision of grades.

Although first year university grade point average has dominated the landscape of predictive validity research, long-term university success is an equally important criterion. The list of recommendations on future studies recommended needs to document admissions criteria on long term university outcomes, such as continuing studies in graduate school, getting hired in career-related occupations, and earning an entry-level salary.

University admissions decisions rely on "hard" and "soft" indicators of academic success in university level studies. Even with the challenges these indicators face, validity studies bring to bear pieces of information to establish basic facts concerning the contribution of PSU scores to predicting success in college. The realization of the challenges faced by university admissions testing programs can help in setting realistic boundaries on what can be expected from elements of admissions criteria. Validation is an everlasting effort to develop lines of supporting evidence on the use and meaning of measures. Several new predictor measures could to be discussed and potentially added in future revisions of Chile's admission criteria. In this context, validity studies are valuable allies for establishing lines of evidence to support decision-making process to move forward with intended changes.

RECOMMENDATIONS

1. We recommend continuing validation efforts for developing lines of supporting evidence on the use and meaning of measures. Several new predictor measures should be added in future revisions of Chile's admission criteria. In this context, we recommend conducting validity studies to establish lines of evidence to support decision-making process to move forward with intended changes.
2. We recommend investigating alternative criteria for predictive validity research beyond first-year university grade point average or graduation rates by including measures of continuing studies in graduate school, of being hired in career-related occupations and of entry-level salary.
3. We recommend investigating whether the university grading practices are uniform within a career at universities and across the same career at different universities, as this information would further ground the findings of predictive validity measures that use university grade point average as a dependent variable

BIBLIOGRAPHY

AERA, APA & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Arce-Ferrer, A., & Borges, I. (2007). Investigating postgraduate college admission interviews: Generalizability Theory reliability and incremental predictive validity. *Journal of Hispanic Higher Education*, *6*(2), 118-134.

Beatty, A., Greenwood, M., & Linn, R. (1999). *Myths and tradeoffs: The role of tests in undergraduate admissions.* National Research Council: The National Academies Press.

Birnbaum, Z., Paulson, E., & Andrews, F. (1950). On the effects of selection performed on some coordinates of a multi-dimensional population*. Psychometrika*, *15* (2), 191-204.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Prediction of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (College Board Report No. 2000-1). New York: College Board.

Burton, N., & Ramist, I. (2001). *Predicting long term success in undergraduate school: A review of predictive validity studies.* (College Board Research Report No. 2001-2). New York: The College Board.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Comité Técnico Asesor. (2010). *Validez diferencial y sesgo de predictividad de las pruebas de admisión a las Universidades Chilenas.* Available at www.cta-psu.cl.

Dunbar, S., & Linn, R. (1991). Range restriction adjustments in the prediction of military job performance. In Beatty, Greenwood and Linn (edits.), *Myth and trade-offs: The role of tests in undergraduate admissions.* National Research Council: The National Academies Press. PP. 127-157.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, New Jersey: Lawrence Erlbaum Associates

Hunter, J., & Schmidt, F. (1990). *Methods of meta-analysis. Correcting errors and bias in research findings.* Sage Publications.

International Test Commission (2000). *International guidelines for test use*. ITC: Author.

Kobrin, J., Patterson, B., Shaw, E., Mattern, K., & Barbuti, S. (2008). *Validity of the SAT for predicting first-year college grade point average*. (College Board Research Report No. 2008-5). New York: The College Board.

Lawley, D. N. A note on Karl Pearson's selection formulae. *Proc. Roy. Soc. Edin., Sect. A. (Math. & Phys.* Sec.),1943-44, 62, Part I, 28-30.I, 28-30.

Linn, R. (1982). Admission testing on trial. *American Psychologist*, *37*, 279-291.

Maxey, J., & Sawyer, R. (1981). *Predictive validity of the ACT assessment for Afro-American/Black, Mexican-American/Chicano, and Caucasian-American/White students* (ACT Research Bulletin 81-1). Iowa City: American College Testing.

MINEDUC. (2012). Descripción metodología de cálculo del ranking del alumno [Memorandum]. Santiago, Chile: Autor.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Research Report No. 93-1). New York: The College Board.

Roberts, W., & Noble, J. (2004). *Academic and Noncognitive variables related to PLAN scores* (ACT Research Report Series 2004-1). Iowa City: ACT, Inc.

Whitney, D. (1989). Educational admissions and placement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 515-525). New York: American Council on Education and Macmillan.

Young, J. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report No. 2001-6). New York: The College Board.

Appendix A. Equating Procedures for the Science Test

2010 Admissions Process
PSU Technical Advisory Committee,
Consejo de Rectores[18]

Introduction

The determining of a single score in the Science test assumes the application of an equating procedure (or more appropriately, linking procedure, using Kolen and Brennan, 2004 terminology). Given the structure of the test, which includes a common section and three alternative modules, the application of a procedure coherent with the test design is required (non-equivalent groups with common questions[19]). The options available in order to establish the score equivalences are several, including linear processes (such as the method by Levine or Tucker), as well as non linear (variations on the equipercentile) or methods based upon IRT.

In the case of university selection tests corresponding to the 2005 admissions process, the option taken was to develop a method which is a variant of the non linear methods[20], the intention is to establish the score equivalences of the optional modules conditioning them to the outputs of the common module. In this procedure, the final score of the Science test is the sum of the common module score (which does not require transformation) plus the adjusted optional module. For this last one, one of the three is used as a reference model, cutting off the solutions to the range of possible values (from -6.5 to 26 points).

Linear models by sections

X is to be the common revised score and y the optional revised score. We wish to express the mean of y in function of x, separately for the Physics, Chemistry and Biology modules. The first part of the process is to adjust linear models by sections, separately for the subpopulations yielded by each model, which may be carried out through a multiple regression routine.

In order to define the sections, m nodes are defined in the x axis, which may or not be equally spaced. In the application to the Science test, the attention is restricted to the values of x between -5 and the maximum score in the common module (54 in the 2006 application, but it may be reduced in case a question is eliminated). It must be considered that the linear model adjustment is only an intermediate step to carry out the equating, in such a way that it is not technically necessary to cover the whole score spectrum.

---

[18] In developing the methodology, the technical advice of Dr. Rianne Jansen, from the Facultad de Educación, of K.U.Leuven, Belgium, was counted upon. She worked jointly with professors Guido del Pino, Jorge Manzi and Ernesto San Martín of the P.U. Católica de Chile., along with the collaboration of Professor Ricardo Aravena, of this same university, concerning information technologies aspects.

[19] It is fitting to assume that the groups are non-equivalent, since the students opt freely between the three modules and it may not be assumed that the three groups are the equivalent of each other.

[20] This approach has been chosen instead of the linear methods, since these latter ones include assumptions which are harder to fulfill with this test. The IRT model, which could be a very appropriate technical solution, has been excluded taking into consideration that the PSU admissions tests, the same as the PAA, are scored using variants of the classical theory.

To discover the procedure we denominate the nodes by a_1, a_2, …, a_m. Through this notation, the adjustment of the linear mode by sections takes place as follows (using a multiple regression routine):

a) Define, for *j*=1, …,*m*, the variables
$Z_j = 0$ if $x < a_j$, $Z_j = x - a_j$, if $a_j < x < a_j+1$, and $Z_j = a_j+1 - a_j$, if $x > a_j+1$

b) Adjustment of the model
$$y(x) = \alpha + \Sigma_J = 1\ \beta_J\ Z_J \qquad (1)$$

to the data, separately for the three modules.

From now on the values for α_y and of the β_k are referred to the estimated values.

Regarding the specific application of the method to the 2010 Science test, the Biology test was used as a basis with the following nodes -13,5; 4; 8; 10; 15; 22; 26; 32; 36; 41; 46; 50; 54.

In the case that a β_k coefficient happened to be negative (a case deemed to be highly improbable) one node must be eliminated and the model must be adjusted again. Naturally, the same would have to be done in the highly hypothetical case in which there should be no score in one of the sections. Regarding the application of the method in the 2010 admissions, there were no negative coefficients with the nodes used finally.

The problem of interpolation

Regarding a given optional score, $y = y_0$ the solution x_0 of the $y(x) = y_0$ equation must be found. For that we defined the constants

$$D_k = a_{k+1} - a_k,\ k=1,\ldots m-1$$
$$\delta_1 = \alpha,\ \ \delta_j = \delta_j + \beta_j D_j,\ \ j=1,\ldots,m-1$$

Depending upon the choice of nodes there may be a solution to this equation.
It exists for $\delta_1 < y0 < \delta_m$. In order to obtain it the procedure is the following:

a) Find the r value which satisfies the δ condition

$r < y_0 < \delta_r+1$

b) Solve the equation

$$\delta_r + \beta_r(x - a_r) = y_0 \qquad (2)$$

Solve the equation
$$x_0 = \alpha_r + (\ y_0 - \delta_r) / \beta_r \qquad (3)$$

The same as in the case of the linear model adjustment, once again interpolation is an intermediate step in order to arrive to the equating. In this sense, it is never necessary to do the interpolation outside the $\delta_1 < y_0 < \delta_m$ range, in such a way that there always is a solution for the equating effects.

The problem of equating

Towards referring ourselves to the reference test we shall use a superindex Ref. Even though the choice of reference test has a marginal effect, the Biology test was used, since it includes a wide spectrum of scores- In such a way that,

$$\delta_{rRef} + \beta_{rRef} (x_{0Ref} - \alpha_r) = y_0 \qquad (4)$$

Substituting (3) in (4) you obtain

$$y_{Ref} (x0) = \delta_{rRef} + \beta_{rRef} (y_0 - \delta_r) / \beta_r \qquad (5)$$

This may be rewritten under the form

$$(y_{Ref} (x0) - \delta_{rRef}) / \beta_{rRef} = (y(x_0) - \delta_r) / \beta_r \qquad (6)$$

which may be interpreted as a simple linear interpolation.

In the especial case $\beta_r = \beta_{rRef}$, the procedure has a simple explanation in terms of the additive connections $\Delta = y_{Ref}(x) - y_0$, which added to the original score y_0 render the equivalent score. In effect, Δ has the form $\Delta = \delta_{rRef} - \delta_r$, that is, the vertical difference in the initial nodes. This is natural since it concerns two parallel segments.

Treatment of the extreme values

The procedure described in the previous section does not deliver directly the additive connections outside the $\delta_1 < y_0 < \delta_m$ range, where these constants are different for the Physics and Chemistry tests, which is why it is necessary to establish how to deal with them. The procedure proposed is the following:

For scores below δ_1, the additive correction calculated for the (-5 to 0) section shall be applied;

For scores above δ_m, the additive correction calculated for the last section shall be applied.

However, the values transformed by the equating procedure must be maintained within the range of possible values, that is, from -6.5 to 26. Values below -6.5 are cut off at -6.5 and those above 26, are cut off at 26. It is worthwhile to remember that the scores transformed of the optional parts must be added to the common score, which tends to attenuate the importance of the extreme values. On the other hand, after applying the normalization process in order to obtain the final values, it is necessary just the same to additionally correct the extreme values in order to limit the standard scores to the established range, between 150 and 850 points. If well enough the treatment of the low scores lacks practical importance, that of the higher scores is more relevant, since it may have influence over those obtaining national scores. This is the main reason for the transformed scores not to exceed 26 points.

Appendix B. Summary Descriptive Item Statistics for Simulated Data Set

Table 166: Summary Descriptive Item Statistics for Simulated Data Set

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT	ITEM*TEST CORRELATION PEARSON	ITEM*TEST CORRELATION BISERIAL
1	ITEM0001	3000.0	1434.0	47.8	0.09	0.236	0.296
2	ITEM0002	3000.0	2767.0	92.2	-2.47	0.339	0.625
3	ITEM0003	3000.0	2137.0	71.2	-0.91	0.411	0.546
4	ITEM0004	3000.0	1368.0	45.6	0.18	0.401	0.504
5	ITEM0005	3000.0	1694.0	56.5	-0.26	0.362	0.456
6	ITEM0006	3000.0	2626.0	87.5	-1.95	0.332	0.534
7	ITEM0007	3000.0	2387.0	79.6	-1.36	0.398	0.566
8	ITEM0008	3000.0	1861.0	62.0	-0.49	0.415	0.529
9	ITEM0009	3000.0	1346.0	44.9	0.21	0.247	0.310
10	ITEM0010	3000.0	1850.0	61.7	-0.48	0.509	0.649
11	ITEM0011	3000.0	1362.0	45.4	0.18	0.401	0.504
12	ITEM0012	3000.0	954.0	31.8	0.76	0.544	0.710
13	ITEM0013	3000.0	1498.0	49.9	0.00	0.366	0.458
14	ITEM0014	3000.0	915.0	30.5	0.82	0.338	0.444
15	ITEM0015	3000.0	1871.0	62.4	-0.51	0.379	0.484
16	ITEM0016	3000.0	2868.0	95.6	-3.08	0.244	0.535
17	ITEM0017	3000.0	1112.0	37.1	0.53	0.381	0.487
18	ITEM0018	3000.0	2479.0	82.6	-1.56	0.400	0.591
19	ITEM0019	3000.0	1161.0	38.7	0.46	0.446	0.567
20	ITEM0020	3000.0	1757.0	58.6	-0.35	0.512	0.647
21	ITEM0021	3000.0	2393.0	79.8	-1.37	0.461	0.657
22	ITEM0022	3000.0	2564.0	85.5	-1.77	0.377	0.581
23	ITEM0023	3000.0	2220.0	74.0	-1.05	0.431	0.582
24	ITEM0024	3000.0	2193.0	73.1	-1.00	0.406	0.546
25	ITEM0025	3000.0	1046.0	34.9	0.62	0.434	0.560
26	ITEM0026	3000.0	1171.0	39.0	0.45	0.455	0.578
27	ITEM0027	3000.0	2186.0	72.9	-0.99	0.493	0.662
28	ITEM0028	3000.0	1690.0	56.3	-0.25	0.406	0.511
29	ITEM0029	3000.0	1867.0	62.2	-0.50	0.414	0.528
30	ITEM0030	3000.0	1794.0	59.8	-0.40	0.362	0.458
31	ITEM0031	3000.0	1291.0	43.0	0.28	0.410	0.517
32	ITEM0032	3000.0	1443.0	48.1	0.08	0.498	0.625
33	ITEM0033	3000.0	1676.0	55.9	-0.24	0.357	0.449
34	ITEM0034	3000.0	1092.0	36.4	0.56	0.286	0.366
35	ITEM0035	3000.0	2455.0	81.8	-1.51	0.407	0.594
36	ITEM0036	3000.0	2130.0	71.0	-0.90	0.335	0.445
37	ITEM0037	3000.0	2891.0	96.4	-3.28	0.246	0.574
38	ITEM0038	3000.0	2390.0	79.7	-1.37	0.457	0.651
39	ITEM0039	3000.0	2868.0	95.6	-3.08	0.239	0.525
40	ITEM0040	3000.0	1463.0	48.8	0.05	0.540	0.677
41	ITEM0041	3000.0	2097.0	69.9	-0.84	0.392	0.516
42	ITEM0042	3000.0	1783.0	59.4	-0.38	0.477	0.603
43	ITEM0043	3000.0	2330.0	77.7	-1.25	0.485	0.676
44	ITEM0044	3000.0	2588.0	86.3	-1.84	0.401	0.628
45	ITEM0045	3000.0	1433.0	47.8	0.09	0.424	0.532
46	ITEM0046	3000.0	2268.0	75.6	-1.13	0.480	0.657
47	ITEM0047	3000.0	1358.0	45.3	0.19	0.166	0.208
48	ITEM0048	3000.0	1079.0	36.0	0.58	0.351	0.451
49	ITEM0049	3000.0	2642.0	88.1	-2.00	0.355	0.578
50	ITEM0050	3000.0	1374.0	45.8	0.17	0.519	0.651

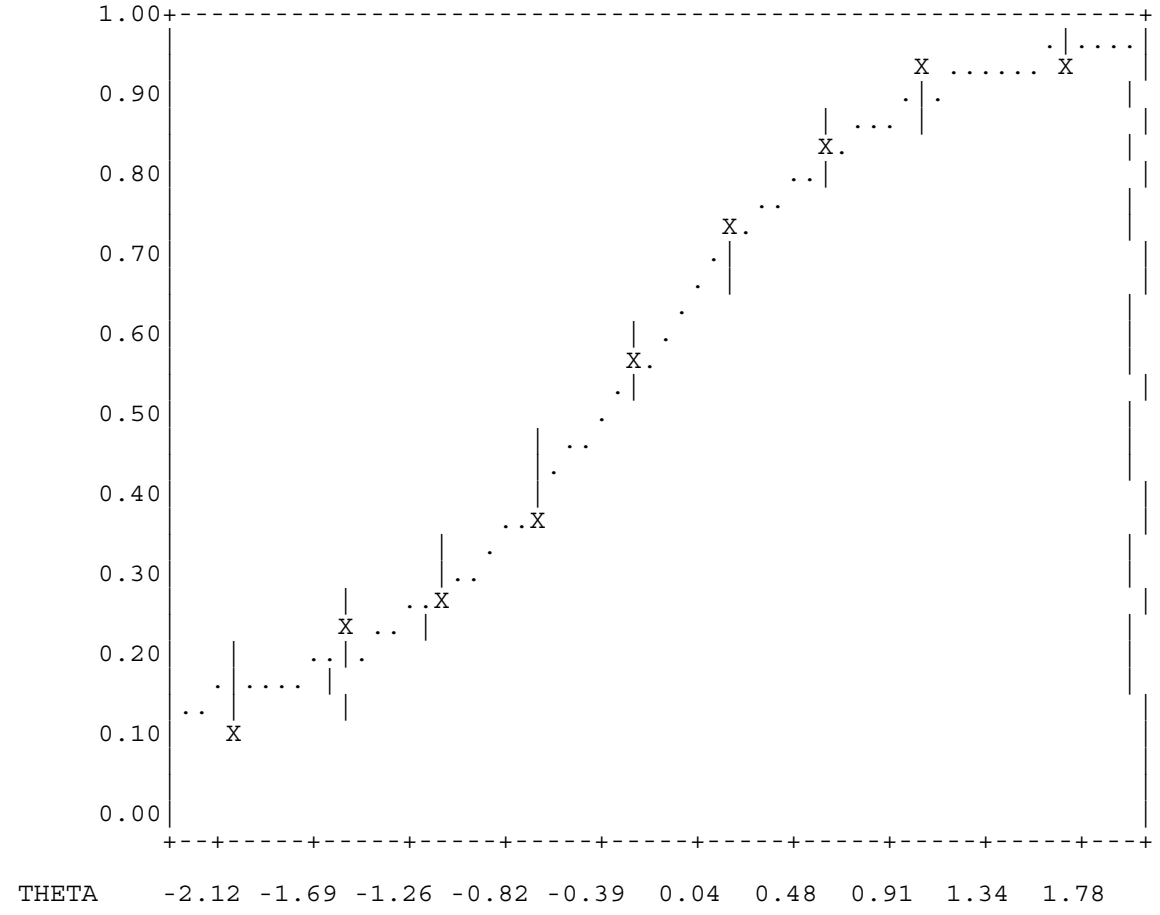


Figure 51: Example of Item Goodness of Fit from Synthetic Data

Table 167: Summary Statistics for Score Estimates from Synthetic Data Set

```
============================================================


 CORRELATIONS AMONG TEST SCORES

                 TEST0001
 TEST0001       1.0000


 MEANS, STANDARD DEVIATIONS, AND VARIANCES OF SCORE ESTIMATES

 TEST:               TEST0001
 MEAN:              -0.0004
 S.D.:               0.9814
 VARIANCE:           0.9631


 ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS

 TEST:               TEST0001
 RMS:               0.2920
 VARIANCE:          0.0853

 EMPIRICAL
   RELIABILITY:     0.9187



 MARGINAL LATENT DISTRIBUTION(S)
 ===============================


 MARGINAL LATENT DISTRIBUTION FOR TEST TEST0001
    MEAN     =     0.000
    S.D.     =     0.998
    SKEWNESS =     0.032
    KURTOSIS =    -0.075

                1           2           3           4           5
 POINT    -0.4000E+01 -0.3111E+01 -0.2222E+01 -0.1333E+01 -0.4444E+00
 WEIGHT    0.9371E-04  0.2755E-02  0.3285E-01  0.1479E+00  0.3259E+00

                6           7           8           9          10
 POINT     0.4444E+00  0.1333E+01  0.2222E+01  0.3111E+01  0.4000E+01
 WEIGHT    0.2986E+00  0.1564E+00  0.3192E-01  0.3457E-02  0.1686E-03
```

Appendix C. Focal and Reference Group Comparisons for Selected Demographic Groups on all Six PSU Tests

Table 168: Frequency of Completed Forms on the Language and Communication Test, by Gender

	Form Number Language		
	Language 101	Language 102	Total
	Frequency	Frequency	Frequency
Gender			
1) Male	53,891	54,034	107,925
2) Female	61,639	61,576	123,215
Total	115,530	115,610	231,140

Table 169: Frequency of Completed Forms on the Mathematics Test, by Gender

	Form Number Mathematics		
	Mathematics 111	Mathematics 112	Total
	Frequency	Frequency	Frequency
Gender			
1) Male	53,891	54,034	107,925
2) Female	61,657	61,558	123,215
Total	115,548	115,592	231,140

Table 170: Frequency of Completed Forms on the History and Social Sciences Test, by Gender

	Form Number History		
	History 121	History 122	Total
	Frequency	Frequency	Frequency
Gender			
1) Male	33,227	32,991	66,218
2) Female	36,940	36,956	73,896
Total	70,167	69,947	140,114

Table 171: Frequency of Completed Forms on the Elective Science Tests, by Gender

	Form Number Elective Science						
	Biology 151	Biology 152	Physics 161	Physics 162	Chemistry 171	Chemistry 172	Total
	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency
Gender							
1) Male	13,533	13,309	10,581	10,338	6,993	6,895	61,649
2) Female	24,636	24,475	3,167	3,057	8,043	7,942	71,320
Total	38,169	37,784	13,748	13,395	15,036	14,837	132,969

Table 172: Frequency of Completed Forms on the Language and Communication Test, by Socioeconomic Status

	Form Number Language		
	Language 101	Language 102	Total
	Frequency	Frequency	Frequency
Socioeconomic Level			
1) Low (1-17 Percentile)	19,534	19,748	39,282
2) Lower Middle (18-52 Percentile)	40,683	40,660	81,343
3) Upper Middle (53-69 Percentile)	19,887	19,519	39,406
4) High (70-83 Percentile)	15,700	15,877	31,577
5) Very High (84-91 Percentile)	8,831	9,046	17,877
6) Top (92-100 Percentile)	10,895	10,760	21,655
Total	115,530	115,610	231,140

Table 173: Frequency of Completed Forms on the Mathematics Test, by Socioeconomic Status

	Form Number Mathematics		
	Mathematics 111	Mathematics 112	Total
	Frequency	Frequency	Frequency
Socioeconomic Level			
1) Low (1-17 Percentile)	19,529	19,753	39,282
2) Lower Middle (18-52 Percentile)	40,698	40,645	81,343
3) Upper Middle (53-69 Percentile)	19,887	19,519	39,406
4) High (70-83 Percentile)	15,708	15,869	31,577
5) Very High (84-91 Percentile)	8,829	9,048	17,877
6) Top (92-100 Percentile)	10,897	10,758	21,655
Total	115,548	115,592	231,140

Table 174: Frequency of Completed Forms on the History and Social Sciences Test, by Socioeconomic Status

	Form Number History		Total
	History 121	History 122	
	Frequency	Frequency	Frequency
Socioeconomic Level			
1) Low (1-17 Percentile)	12,880	12,724	25,604
2) Lower Middle (18-52 Percentile)	25,825	25,587	51,412
3) Upper Middle (53-69 Percentile)	11,684	11,870	23,554
4) High (70-83 Percentile)	8,812	9,040	17,852
5) Very High (84-91 Percentile)	4,910	4,733	9,643
6) Top (92-100 Percentile)	6,056	5,993	12,049
Total	70,167	69,947	140,114

Table 175: Frequency of Completed Forms on the Elective Science Tests, by Socioeconomic Status

	Form Number Elective Science						Total
	Biology 151	Biology 152	Physics 161	Physics 162	Chemistry 171	Chemistry 172	
	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency
Socioeconomic Level							
1) Low (1-17 Percentile)	7,121	7,238	1,644	1,637	1,845	1,879	21,364
2) Lower Middle (18-52 Percentile)	13,952	13,732	3,997	3,804	4,437	4,402	44,324
3) Upper Middle (53-69 Percentile)	6,440	6,373	2,278	2,254	2,640	2,601	22,586
4) High (70-83 Percentile)	5,150	5,073	2,151	2,171	2,416	2,431	19,392
5) Very High (84-91 Percentile)	2,762	2,739	1,486	1,349	1,639	1,534	11,509
6) Top (92-100 Percentile)	2,744	2,629	2,192	2,180	2,059	1,990	13,794
Total	38,169	37,784	13,748	13,395	15,036	14,837	132,969

Table 176: Frequency of Completed Forms on the Language and Communication Test, by Type of High School

	Form Number Language		
	Language 101	Language 102	Total
	Frequency	Frequency	Frequency
Curricular Branch			
1) Scientific-Humanistic	82,505	82,565	165,070
2) Technical-Professional	33,025	33,045	66,070
Total	115,530	115,610	231,140

Table 177: Frequency of Completed Forms on the Mathematics Test, by Type of High School

	Form Number Mathematics		
	Mathematics 111	Mathematics 112	Total
	Frequency	Frequency	Frequency
Curricular Branch			
1) Scientific-Humanistic	82,515	82,555	165,070
2) Technical-Professional	33,033	33,037	66,070
Total	115,548	115,592	231,140

Table 178: Frequency of Completed Forms on the History and Social Sciences Test, by Type of High School

	Form Number History		
	History 121	History 122	Total
	Frequency	Frequency	Frequency
Curricular Branch			
1) Scientific-Humanistic	46,908	46,419	93,327
2) Technical-Professional	23,259	23,528	46,787
Total	70,167	69,947	140,114

Table 179: Frequency of Completed Forms on the Elective Science Tests, by Type of High School

	Form Number Elective Science						
	Biology 151	Biology 152	Physics 161	Physics 162	Chemistry 171	Chemistry 172	Total
	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency
Curricular Branch							
1) Scientific-Humanistic	29,228	28,922	10,664	10,391	13,080	12,872	105,157
2) Technical-Professional	8,941	8,862	3,084	3,004	1,956	1,965	27,812
Total	38,169	37,784	13,748	13,395	15,036	14,837	132,969

Table 180: Frequency of Completed Forms on the Language and Communication Test, by Region

	Form Number Language		
	Language 101	Language 102	Total
	Frequency	Frequency	Frequency
Region			
1) North	12,970	13,016	25,986
2) Central	59,726	59,709	119,435
3) South	42,834	42,885	85,719
Total	115,530	115,610	231,140

Table 181: Frequency of Completed Forms on the Mathematics Test, by Region

	Form Number Mathematics		
	Mathematics 111	Mathematics 112	Total
	Frequency	Frequency	Frequency
Region			
1) North	12,962	13,024	25,986
2) Central	59,741	59,694	119,435
3) South	42,845	42,874	85,719
Total	115,548	115,592	231,140

Table 182: Frequency of Completed Forms on the History and Social Sciences Test, by Region

	Form Number History		
	History 121	History 122	Total
	Frequency	Frequency	Frequency
Region			
1) North	7,561	7,605	15,166
2) Central	36,883	36,706	73,589
3) South	25,723	25,636	51,359
Total	70,167	69,947	140,114

Table 183: Frequency of Completed Forms on the Elective Science Tests, by Region

	Form Number Elective Science						
	Biology 151	Biology 152	Physics 161	Physics 162	Chemistry 171	Chemistry 172	Total
	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency
Region							
1) North	3,994	3,905	1,785	1,733	2,204	2,171	15,792
2) Central	17,469	17,205	7,246	7,019	7,047	6,995	62,981
3) South	16,706	16,674	4,717	4,643	5,785	5,671	54,196
Total	38,169	37,784	13,748	13,395	15,036	14,837	132,969

Table 184: Frequency of Completed Forms on the Language and Communication Test, by High School Financing

	Form Number Language		Total
	Language 101	Language 102	
	Frequency	Frequency	Frequency
High School Financing			
1) Private	12,354	12,193	24,547
2) Subsidized	60,323	60,847	121,170
3) Municipal	41,764	41,467	83,231
Total	114,441	114,507	228,948

Table 185: Frequency of Completed Forms on the Mathematics Test, by High School Financing

	Form Number Mathematics		Total
	Mathematics 111	Mathematics 112	
	Frequency	Frequency	Frequency
High School Financing			
1) Private	12,345	12,202	24,547
2) Subsidized	60,335	60,835	121,170
3) Municipal	41,774	41,457	83,231
Total	114,454	114,494	228,948

Table 186: Frequency of Completed Forms on the History and Social Sciences Test, by High School Financing

	Form Number History		Total
	History 121	History 122	
	Frequency	Frequency	Frequency
High School Financing			
1) Private	6,990	6,925	13,915
2) Subsidized	36,031	36,039	72,070
3) Municipal	26,507	26,325	52,832
Total	69,528	69,289	138,817

Table 187: Frequency of Completed Forms on the Elective Science Tests, by High School Financing

	Form Number Elective Science						Total
	Biology 151	Biology 152	Physics 161	Physics 162	Chemistry 171	Chemistry 172	
	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency	Frequency
High School Financing							
1) Private	3,096	3,007	2,251	2,243	2,235	2,205	15,037
2) Subsidized	20,310	20,216	7,212	6,938	8,222	8,061	70,959
3) Municipal	14,362	14,181	4,171	4,107	4,472	4,470	45,763
Total	37,768	37,404	13,634	13,288	14,929	14,736	131,759

Appendix D. Fitting a Three-Parameter Logistic Function to the Response

Table 188: Factor Analysis Communalities, CTT Difficulty (*p*) and Discrimination (*r*), IRT Log of Discrimination (*log a*), Difficulty *b*, Guessing *c*, Goodness of Fit Chi-Square and Probability, and Root Mean Square Error (*RMSE*) for the Language and Communication Test

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
L01	0.207	0.859	0.124	-1.596	-5.021	0.093	190.061	0.000	0.020
L02	0.498	0.739	0.296	-0.747	-1.398	0.008	259.598	0.000	0.029
L03	0.696	0.849	0.380	-0.135	-1.453	0.020	53.402	0.000	0.009
L04	0.635	0.750	0.408	-0.315	-0.927	0.095	166.299	0.000	0.020
L05	0.330	0.670	0.311	-0.139	0.483	0.466	56.682	0.000	0.016
L06	0.342	0.637	0.289	-0.046	0.643	0.451	17.440	0.000	0.009
L07	0.516	0.657	0.320	-0.777	-0.746	0.047	101.919	0.000	0.020
L08	0.518	0.629	0.401	-0.187	0.082	0.244	10.334	0.001	0.006
L09	0.360	0.574	0.291	-0.704	0.320	0.226	2.207	0.137	0.003
L10	0.729	0.544	0.565	0.394	0.213	0.143	1.821	0.177	0.002
L11	0.543	0.592	0.367	-0.506	0.025	0.110	123.393	0.000	0.023
L12	0.486	0.483	0.441	-0.009	0.479	0.178	4.237	0.040	0.004
L13	0.579	0.474	0.406	-0.420	0.357	0.028	32.817	0.000	0.011
L14	0.553	0.438	0.499	0.156	0.549	0.127	43.101	0.000	0.010
L15	0.600	0.370	0.462	-0.178	0.699	0.038	140.829	0.000	0.025
L16	0.526	0.899	0.304	-0.117	-1.392	0.433	14.503	0.000	0.005
L17	0.534	0.736	0.397	-0.223	-0.596	0.236	13.443	0.000	0.006
L18	0.567	0.635	0.428	-0.120	-0.082	0.217	19.159	0.000	0.006
L19	0.426	0.458	0.353	-0.460	0.699	0.149	17.932	0.000	0.008
L20	0.689	0.333	0.567	0.251	0.673	0.031	36.280	0.000	0.009
L21	0.669	0.815	0.423	-0.005	-1.088	0.133	8.653	0.003	0.003
L22	0.441	0.783	0.337	0.070	-0.070	0.533	11.036	0.001	0.006
L23	0.696	0.790	0.458	0.171	-0.677	0.239	8.558	0.003	0.005
L24	0.280	0.676	0.215	-1.053	-0.290	0.286	49.006	0.000	0.015
L25	0.487	0.751	0.314	-0.678	-1.300	0.028	34.469	0.000	0.010
L26	0.488	0.600	0.397	-0.307	0.051	0.187	1.351	0.245	0.002
L27	0.607	0.715	0.446	-0.292	-0.614	0.109	13.059	0.000	0.007
L28	0.395	0.503	0.286	-0.963	0.191	0.030	7.247	0.007	0.006
L29	0.613	0.408	0.525	0.326	0.624	0.120	85.527	0.000	0.017
L30	0.154	0.252	0.115	-1.886	5.536	0.065	18.120	0.000	0.009
L31	0.658	0.848	0.383	-0.056	-1.212	0.194	11.537	0.001	0.003
L32	0.692	0.803	0.413	-0.128	-1.241	0.005	220.406	0.000	0.019
L33	0.694	0.848	0.390	-0.018	-1.085	0.259	71.582	0.000	0.011
L34	0.655	0.853	0.366	-0.235	-1.580	0.039	10.504	0.001	0.004
L35	0.591	0.653	0.451	-0.108	-0.233	0.193	1.801	0.180	0.002
L36	0.410	0.659	0.298	-0.756	-0.506	0.169	3.950	0.047	0.004
L37	0.414	0.496	0.275	-1.016	0.128	0.016	75.797	0.000	0.018

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
L38	0.322	0.817	0.188	-1.146	-2.868	0.029	165.648	0.000	0.022
L39	0.528	0.709	0.410	-0.106	-0.259	0.259	11.796	0.001	0.005
L41	0.540	0.726	0.380	-0.472	-1.037	0.044	5.540	0.019	0.004
L42	0.371	0.808	0.259	-0.298	-0.241	0.558	10.934	0.001	0.007
L43	0.430	0.790	0.244	-0.939	-2.042	0.028	52.960	0.000	0.014
L44	0.445	0.419	0.308	-0.895	0.820	0.053	46.014	0.000	0.017
L45	0.470	0.669	0.341	-0.729	-0.871	0.010	90.551	0.000	0.018
L46	0.392	0.613	0.259	-1.027	-0.533	0.070	10.700	0.001	0.008
L47	0.576	0.683	0.452	0.328	0.170	0.349	46.879	0.000	0.016
L48	0.393	0.425	0.277	-1.093	0.755	0.015	63.901	0.000	0.022
L49	0.314	0.414	0.277	-0.829	1.058	0.125	9.784	0.002	0.006
L50	0.329	0.845	0.215	-0.665	-0.937	0.527	65.564	0.000	0.014
L51	0.411	0.736	0.251	-0.890	-1.544	0.007	616.906	0.000	0.044
L52	0.329	0.295	0.275	-0.170	1.508	0.150	17.353	0.000	0.008
L53	0.357	0.114	0.209	-0.910	3.260	0.002	102.131	0.000	0.016
L54	0.511	0.591	0.354	-0.673	-0.345	0.058	52.447	0.000	0.013
L55	0.631	0.906	0.330	-0.014	-1.784	0.061	11.518	0.001	0.004
L56	0.691	0.711	0.433	-0.278	-0.783	0.012	57.225	0.000	0.012
L57	0.326	0.719	0.217	-1.166	-1.684	0.051	22.772	0.000	0.009
L58	0.428	0.366	0.340	-0.494	1.099	0.101	78.406	0.000	0.018
L59	0.333	0.364	0.325	0.004	1.310	0.192	68.597	0.000	0.019
L60	0.345	0.262	0.235	-1.056	2.031	0.009	163.917	0.000	0.025
L61	0.587	0.592	0.476	0.036	0.092	0.197	3.528	0.060	0.004
L62	0.227	0.319	0.206	-0.790	2.152	0.158	20.663	0.000	0.009
L63	0.259	0.535	0.223	-1.261	-0.225	0.017	53.259	0.000	0.015
L64	0.566	0.615	0.440	0.050	0.296	0.251	20.703	0.000	0.009
L66	0.349	0.610	0.235	-1.170	-0.668	0.014	95.683	0.000	0.022
L67	0.762	0.694	0.470	-0.269	-0.649	0.007	170.367	0.000	0.027
L68	0.367	0.783	0.258	-0.882	-1.746	0.124	6.705	0.010	0.005
L69	0.579	0.931	0.273	-0.173	-2.147	0.207	3.095	0.079	0.002
L70	0.604	0.592	0.451	-0.255	-0.112	0.085	22.613	0.000	0.008
L71	0.551	0.356	0.407	-0.152	0.920	0.100	12.680	0.000	0.007
L72	0.378	0.459	0.319	-0.586	0.714	0.150	65.770	0.000	0.016
L73	0.329	0.640	0.278	-0.894	-0.591	0.120	7.043	0.008	0.004
L74	0.693	0.405	0.545	-0.040	0.530	0.044	68.962	0.000	0.018
L75	0.812	0.817	0.526	0.384	-0.776	0.103	3.834	0.050	0.003
L76	0.554	0.530	0.493	0.022	0.236	0.110	2.412	0.120	0.003
L77	0.621	0.761	0.424	0.010	-0.480	0.264	10.382	0.001	0.004
L78	0.495	0.519	0.424	-0.145	0.494	0.162	29.072	0.000	0.010
L79	0.751	0.629	0.571	0.221	-0.212	0.071	3.128	0.077	0.002
L80	0.197	0.545	0.154	-1.908	-0.377	0.041	126.795	0.000	0.027

Table 189: Factor Analysis Communalities, CTT Difficulty (*p*) and Discrimination (*r*), IRT Log of Discrimination (*log a*), Difficulty *b*, Guessing *c*, Goodness of Fit Chi-Square and Probability, and Root Mean Square Error (*RMSE*) for the Mathematics Test

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
M01	0.723	0.788	0.474	0.368	-0.463	0.299	7.765	0.005	0.003
M02	0.612	0.826	0.359	-0.050	-1.009	0.276	10.529	0.001	0.004
M03	0.780	0.781	0.480	0.308	-0.802	0.069	10.995	0.001	0.005
M04	0.565	0.763	0.410	-0.061	-0.471	0.306	12.030	0.001	0.006
M05	0.758	0.742	0.547	0.646	-0.293	0.229	30.333	0.000	0.008
M06	0.685	0.779	0.395	-0.110	-1.113	0.006	93.312	0.000	0.014
M07	0.732	0.654	0.574	0.402	-0.269	0.109	47.418	0.000	0.009
M08	0.739	0.752	0.547	0.500	0.038	0.259	3.941	0.047	0.003
M09	0.618	0.372	0.568	0.794	0.961	0.123	136.798	0.000	0.024
M10	0.674	0.176	0.566	0.983	1.273	0.034	136.923	0.000	0.014
M11	0.809	0.853	0.445	0.513	-0.952	0.010	113.921	0.000	0.013
M12	0.716	0.783	0.468	0.326	-0.581	0.207	36.402	0.000	0.007
M13	0.614	0.747	0.388	-0.254	-1.051	0.015	28.479	0.000	0.009
M14	0.833	0.733	0.573	0.459	-0.319	0.105	56.435	0.000	0.015
M15	0.657	0.640	0.473	-0.013	-0.121	0.130	93.711	0.000	0.018
M16	0.857	0.774	0.573	0.596	-0.341	0.203	31.870	0.000	0.010
M17	0.913	0.763	0.634	0.594	-0.263	0.097	56.846	0.000	0.010
M18	0.773	0.662	0.572	0.400	-0.149	0.144	100.370	0.000	0.016
M19	0.593	0.680	0.452	0.143	0.025	0.313	45.463	0.000	0.013
M20	0.731	0.620	0.598	0.551	0.257	0.171	8.199	0.004	0.005
M21	0.878	0.573	0.723	0.875	0.222	0.078	39.847	0.000	0.010
M22	0.647	0.747	0.446	0.002	-0.809	0.008	221.471	0.000	0.022
M23	0.791	0.626	0.614	0.560	0.076	0.159	19.321	0.000	0.007
M24	0.651	0.551	0.517	0.200	0.346	0.174	341.268	0.000	0.035
M25	0.815	0.648	0.656	0.734	0.429	0.128	65.703	0.000	0.017
M26	0.828	0.785	0.522	0.350	-0.676	0.092	155.554	0.000	0.017
M27	0.767	0.759	0.567	0.421	-0.163	0.183	48.222	0.000	0.012
M28	0.482	0.534	0.378	-0.529	-0.019	0.051	25.238	0.000	0.008
M29	0.826	0.814	0.569	0.505	-0.357	0.163	60.065	0.000	0.016
M30	0.622	0.767	0.432	0.389	0.163	0.437	25.568	0.000	0.010
M31	0.631	0.321	0.555	1.002	1.043	0.102	122.264	0.000	0.019
M32	0.763	0.744	0.557	0.678	0.116	0.308	144.964	0.000	0.018
M33	0.800	0.660	0.608	0.312	0.182	0.187	18.906	0.000	0.009
M34	0.684	0.585	0.539	0.857	0.660	0.298	157.416	0.000	0.033
M35	0.627	0.728	0.423	-0.116	-0.511	0.220	38.713	0.000	0.010
M36	0.613	0.726	0.499	0.657	0.405	0.394	15.578	0.000	0.010
M37	0.483	0.476	0.399	0.095	0.997	0.211	51.666	0.000	0.018
M38	0.775	0.359	0.661	0.948	1.000	0.086	76.612	0.000	0.020
M39	0.560	0.636	0.458	0.028	0.084	0.256	67.543	0.000	0.017

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
M40	0.477	0.526	0.462	0.784	0.755	0.326	36.863	0.000	0.013
M41	0.732	0.600	0.589	0.464	0.308	0.169	159.249	0.000	0.024
M42	0.678	0.352	0.627	0.727	0.775	0.079	388.588	0.000	0.027
M43	0.635	0.510	0.516	0.230	0.579	0.184	82.871	0.000	0.019
M44	0.702	0.438	0.555	0.069	0.710	0.081	111.909	0.000	0.026
M45	0.824	0.713	0.610	0.613	0.043	0.174	8.123	0.004	0.006
M46	0.531	0.766	0.373	-0.430	-1.076	0.082	161.164	0.000	0.022
M47	0.673	0.556	0.549	0.457	0.538	0.202	6.721	0.010	0.006
M48	0.914	0.424	0.771	0.652	0.576	0.035	394.455	0.000	0.044
M49	0.790	0.550	0.696	0.396	0.509	0.148	31.074	0.000	0.022
M50	0.502	0.263	0.495	1.011	1.320	0.099	141.803	0.000	0.028
M51	0.918	0.731	0.685	0.679	-0.146	0.112	21.027	0.000	0.008
M52	0.820	0.636	0.629	0.677	0.397	0.218	115.520	0.000	0.025
M53	0.869	0.506	0.664	0.658	0.684	0.105	69.726	0.000	0.019
M54	0.371	0.263	0.418	0.800	1.326	0.117	60.393	0.000	0.016
M55	0.723	0.682	0.547	0.611	0.131	0.311	13.131	0.000	0.005
M56	0.865	0.639	0.658	0.453	0.210	0.105	41.171	0.000	0.015
M57	0.570	0.363	0.469	0.587	1.199	0.152	10.628	0.001	0.009
M58	0.718	0.370	0.572	1.017	1.078	0.094	162.283	0.000	0.025
M59	0.744	0.787	0.449	0.170	-0.974	0.028	51.615	0.000	0.009
M60	0.643	0.729	0.434	-0.050	-0.627	0.129	23.445	0.000	0.008
M61	0.609	0.531	0.545	0.521	0.534	0.221	138.840	0.000	0.024
M62	0.352	0.246	0.424	0.601	1.315	0.111	378.538	0.000	0.034
M63	0.412	0.133	0.353	0.161	1.815	0.042	96.417	0.000	0.016
M64	0.529	0.510	0.361	-0.681	0.151	0.030	40.495	0.000	0.015
M65	0.658	0.216	0.588	0.730	1.133	0.039	105.375	0.000	0.012
M66	0.594	0.759	0.399	-0.053	-0.584	0.257	19.616	0.000	0.006
M67	0.586	0.633	0.458	0.064	-0.001	0.235	21.666	0.000	0.007
M68	0.525	0.259	0.504	0.669	1.120	0.094	266.780	0.000	0.026
M70	0.620	0.694	0.469	0.039	-0.324	0.187	51.955	0.000	0.011
M71	0.749	0.592	0.616	0.620	0.301	0.186	77.337	0.000	0.017
M72	0.620	0.499	0.525	0.312	0.637	0.208	13.447	0.000	0.006
M73	0.638	0.461	0.499	0.391	0.820	0.186	26.540	0.000	0.009
M74	0.608	0.455	0.551	0.703	0.801	0.180	242.044	0.000	0.033
M75	0.702	0.735	0.429	-0.307	-0.750	0.005	999.742	0.000	0.063

Table 190: Factor Analysis Communalities, CTT Difficulty (*p*) and Discrimination (*r*), IRT Log of Discrimination (*log a*), Difficulty *b*, Guessing *c*, Goodness of Fit Chi-Square and Probability, and Root Mean Square Error (*RMSE*) for the History and Social Sciences Test

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
H01	0.667	0.583	0.570	0.245	-0.062	0.105	30.675	0.000	0.011
H02	0.672	0.651	0.538	0.073	-0.263	0.108	13.308	0.000	0.009
H03	0.491	0.624	0.385	-0.320	0.010	0.219	9.212	0.002	0.008
H04	0.797	0.536	0.567	0.034	0.150	0.049	113.099	0.000	0.029
H05	0.599	0.866	0.367	0.094	-1.101	0.329	7.345	0.007	0.005
H06	0.671	0.335	0.563	0.534	0.754	0.086	74.511	0.000	0.017
H07	0.496	0.854	0.326	-0.327	-1.677	0.054	16.556	0.000	0.007
H08	0.484	0.755	0.324	-0.590	-1.339	0.029	19.250	0.000	0.010
H09	0.667	0.487	0.488	-0.218	0.086	0.010	86.204	0.000	0.023
H10	0.534	0.662	0.422	-0.247	-0.281	0.170	11.428	0.001	0.009
H11	0.514	0.775	0.328	-0.545	-1.321	0.096	10.134	0.001	0.007
H12	0.476	0.686	0.362	-0.411	-0.309	0.255	2.035	0.154	0.003
H13	0.629	0.465	0.553	0.208	0.354	0.110	7.257	0.007	0.005
H14	0.785	0.799	0.523	0.306	-0.648	0.202	9.806	0.002	0.008
H15	0.434	0.786	0.322	-0.531	-1.223	0.193	12.429	0.000	0.007
H16	0.653	0.627	0.489	0.066	-0.054	0.168	5.245	0.022	0.005
H17	0.711	0.731	0.504	0.097	-0.464	0.168	2.830	0.093	0.002
H18	0.521	0.742	0.350	-0.403	-0.659	0.222	13.982	0.000	0.008
H19	0.364	0.482	0.367	-0.559	0.367	0.102	35.269	0.000	0.016
H20	0.708	0.790	0.460	0.016	-0.710	0.173	29.386	0.000	0.010
H21	0.754	0.645	0.578	0.324	-0.125	0.115	11.986	0.001	0.007
H22	0.661	0.815	0.405	-0.096	-1.303	0.016	66.894	0.000	0.012
H23	0.645	0.267	0.498	0.036	1.000	0.040	79.359	0.000	0.018
H24	0.503	0.759	0.340	-0.474	-1.077	0.111	5.087	0.024	0.005
H25	0.304	0.636	0.242	-1.125	-0.896	0.053	22.182	0.000	0.013
H26	0.571	0.758	0.417	-0.043	-0.531	0.307	5.558	0.018	0.005
H27	0.710	0.672	0.566	0.638	-0.069	0.237	13.644	0.000	0.009
H28	0.581	0.554	0.461	-0.001	0.352	0.199	26.257	0.000	0.012
H29	0.548	0.443	0.441	-0.169	0.580	0.121	9.220	0.002	0.008
H30	0.803	0.471	0.628	0.423	0.378	0.081	33.511	0.000	0.013
H31	0.620	0.586	0.470	-0.235	-0.115	0.087	64.819	0.000	0.023
H32	0.584	0.754	0.427	0.021	-0.281	0.330	11.924	0.001	0.008
H33	0.768	0.674	0.543	0.245	-0.106	0.165	35.124	0.000	0.015
H34	0.694	0.789	0.489	0.107	-0.801	0.121	50.869	0.000	0.015
H35	0.569	0.663	0.483	0.004	-0.331	0.174	1.729	0.189	0.003
H36	0.581	0.823	0.406	0.317	-0.504	0.461	14.077	0.000	0.007
H37	0.623	0.745	0.448	0.038	-0.513	0.236	20.280	0.000	0.010
H38	0.714	0.538	0.577	0.518	0.387	0.163	8.492	0.004	0.008
H39	0.556	0.559	0.413	0.517	0.803	0.351	27.975	0.000	0.020

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
H40	0.709	0.771	0.512	0.291	-0.531	0.239	6.059	0.014	0.005
H41	0.646	0.799	0.445	0.052	-0.911	0.140	26.140	0.000	0.006
H42	0.552	0.681	0.400	-0.323	-0.278	0.212	22.911	0.000	0.015
H43	0.671	0.747	0.460	0.020	-0.651	0.163	17.157	0.000	0.008
H44	0.728	0.787	0.495	0.283	-0.626	0.226	12.056	0.001	0.005
H45	0.699	0.351	0.525	0.609	0.917	0.088	126.294	0.000	0.026
H46	0.647	0.794	0.393	-0.118	-0.782	0.216	16.405	0.000	0.007
H47	0.551	0.683	0.429	-0.036	-0.253	0.232	5.788	0.016	0.005
H48	0.690	0.732	0.461	-0.111	-0.694	0.061	14.907	0.000	0.010
H49	0.591	0.779	0.410	-0.134	-0.906	0.191	49.456	0.000	0.012
H50	0.540	0.726	0.416	-0.169	-0.468	0.179	7.736	0.005	0.007
H51	0.822	0.720	0.529	0.079	-0.311	0.157	8.535	0.003	0.008
H52	0.566	0.383	0.439	0.144	0.883	0.154	51.084	0.000	0.020
H53	0.562	0.628	0.443	-0.259	-0.281	0.093	13.074	0.000	0.009
H54	0.676	0.754	0.453	-0.114	-0.787	0.045	21.921	0.000	0.010
H55	0.785	0.530	0.637	0.472	0.160	0.079	11.263	0.001	0.006
H56	0.680	0.729	0.472	0.111	-0.437	0.213	18.498	0.000	0.011
H57	0.544	0.723	0.412	-0.273	-0.402	0.195	10.357	0.001	0.008
H58	0.664	0.727	0.495	0.132	-0.450	0.173	4.196	0.041	0.005
H59	0.707	0.661	0.523	0.066	-0.311	0.085	94.545	0.000	0.019
H60	0.567	0.750	0.407	-0.266	-0.812	0.030	46.321	0.000	0.018
H61	0.605	0.704	0.461	-0.068	-0.553	0.147	11.148	0.001	0.008
H62	0.718	0.559	0.541	0.164	0.124	0.110	31.318	0.000	0.015
H63	0.717	0.794	0.445	0.191	-0.585	0.247	11.049	0.001	0.008
H64	0.677	0.562	0.522	0.190	0.235	0.164	14.439	0.000	0.010
H65	0.677	0.634	0.524	0.233	0.080	0.198	20.289	0.000	0.012
H66	0.655	0.729	0.517	0.053	-0.450	0.167	16.018	0.000	0.010
H67	0.714	0.757	0.516	0.375	-0.533	0.223	10.319	0.001	0.005
H68	0.727	0.544	0.583	0.235	0.132	0.097	30.363	0.000	0.013
H69	0.692	0.818	0.436	-0.061	-0.922	0.171	4.478	0.034	0.005
H70	0.622	0.646	0.472	-0.042	-0.164	0.195	3.274	0.070	0.004
H71	0.747	0.730	0.530	0.202	-0.315	0.198	3.698	0.054	0.005
H72	0.748	0.722	0.528	0.321	-0.344	0.204	9.164	0.002	0.005
H73	0.659	0.573	0.543	0.251	0.132	0.147	26.542	0.000	0.012
H74	0.802	0.695	0.567	0.248	-0.168	0.111	20.585	0.000	0.012
H75	0.558	0.530	0.441	0.032	0.437	0.228	12.356	0.000	0.008

Table 191: Factor Analysis Communalities, CTT Difficulty (*p*) and Discrimination (*r*), IRT Log of Discrimination (*log a*), Difficulty *b*, Guessing *c*, Goodness of Fit Chi-Square and Probability, and Root Mean Square Error (*RMSE*) for the Science - Biology Test

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
B02C	0.703	0.630	0.438	0.320	-0.005	0.214	0.997	0.318	0.003
B03C	0.827	0.399	0.506	0.097	0.466	0.022	19.423	0.000	0.012
B07C	0.437	0.508	0.325	0.045	0.801	0.258	0.554	0.457	0.002
B13C	0.625	0.403	0.410	-0.066	0.521	0.054	4.522	0.033	0.005
B14C	0.705	0.521	0.428	0.078	0.219	0.103	1.478	0.224	0.003
B16C	0.641	0.717	0.344	0.088	-0.521	0.173	0.249	0.618	0.001
B20C	0.611	0.753	0.286	-0.145	-0.762	0.039	3.419	0.064	0.004
B23C	0.621	0.455	0.408	0.213	0.735	0.191	0.865	0.352	0.003
B25C	0.700	0.565	0.476	0.448	0.190	0.168	2.575	0.109	0.004
B26C	0.857	0.533	0.484	0.395	0.189	0.062	2.171	0.141	0.003
B28C	0.390	0.715	0.207	-0.744	-1.078	0.092	3.211	0.073	0.004
B30C	0.748	0.554	0.434	0.121	0.061	0.053	15.910	0.000	0.009
B32C	0.619	0.662	0.351	0.018	-0.203	0.181	3.305	0.069	0.003
B34C	0.766	0.747	0.360	0.367	-0.239	0.233	8.454	0.004	0.007
B38C	0.392	0.580	0.303	0.390	0.609	0.376	3.917	0.048	0.006
B40C	0.531	0.478	0.335	-0.238	0.564	0.149	39.927	0.000	0.019
B44C	0.657	0.632	0.318	0.012	-0.178	0.070	2.520	0.112	0.004
B01	0.401	0.716	0.330	-0.529	-0.719	0.219	5.082	0.024	0.007
B04	0.448	0.416	0.396	0.201	1.077	0.250	2.295	0.130	0.006
B05	0.496	0.522	0.360	-0.327	0.800	0.282	12.075	0.001	0.016
B06	0.297	0.362	0.265	-0.440	1.718	0.230	29.774	0.000	0.022
B08	0.700	0.363	0.520	0.458	0.918	0.154	7.011	0.008	0.010
B09	0.499	0.620	0.395	-0.430	-0.174	0.187	0.906	0.341	0.004
B11	0.702	0.444	0.490	-0.150	0.611	0.127	2.641	0.104	0.008
B12	0.759	0.423	0.590	0.905	0.610	0.187	28.695	0.000	0.018
B15	0.633	0.360	0.546	0.460	0.721	0.142	1.733	0.188	0.005
B17	0.667	0.395	0.512	0.206	0.913	0.135	39.438	0.000	0.030
B18	0.455	0.181	0.345	-0.214	1.689	0.056	25.088	0.000	0.018
B19	0.499	0.764	0.379	-0.305	-1.161	0.102	8.557	0.003	0.007
B21	0.519	0.238	0.417	0.138	1.322	0.099	29.830	0.000	0.019
B22	0.641	0.762	0.434	0.176	-0.631	0.230	3.533	0.060	0.004
B24	0.500	0.445	0.465	-0.020	0.555	0.163	2.930	0.087	0.006
B27	0.586	0.249	0.479	0.237	1.176	0.087	7.805	0.005	0.009
B29	0.752	0.546	0.550	0.069	-0.044	0.061	2.379	0.123	0.006
B31	0.676	0.442	0.537	-0.094	0.101	0.010	38.654	0.000	0.023
B33	0.528	0.543	0.404	-0.464	0.031	0.095	6.814	0.009	0.011
B35	0.656	0.367	0.529	-0.105	0.508	0.032	36.891	0.000	0.021
B36	0.706	0.368	0.544	0.116	0.751	0.075	10.293	0.001	0.011
B37	0.110	0.330	0.182	-1.488	2.984	0.107	11.732	0.001	0.016

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
B39	0.441	0.737	0.317	-0.590	-0.916	0.166	6.591	0.010	0.010
B41	0.744	0.364	0.576	0.274	0.562	0.085	17.366	0.000	0.013
B42	0.338	0.456	0.319	-0.109	1.215	0.300	4.752	0.029	0.010
B43	0.620	0.715	0.461	0.078	-0.579	0.163	3.184	0.074	0.005

Table 192: Factor Analysis Communalities, CTT Difficulty (*p*) and Discrimination (*r*), IRT Log of Discrimination (*log a*), Difficulty *b*, Guessing *c*, Goodness of Fit Chi-Square and Probability, and Root Mean Square Error (*RMSE*) for the Science - Physics Test

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
F02C	0.673	0.647	0.393	0.019	-0.222	0.137	3.456	0.063	0.004
F03C	0.658	0.783	0.382	0.456	-0.595	0.266	0.838	0.360	0.002
F04C	0.561	0.675	0.393	0.065	-0.040	0.215	9.984	0.002	0.009
F07C	0.393	0.743	0.305	-0.517	-1.059	0.038	21.558	0.000	0.012
F11C	0.700	0.595	0.473	0.378	0.174	0.160	9.561	0.002	0.008
F12C	0.540	0.797	0.304	-0.181	-1.194	0.016	32.674	0.000	0.013
F15C	0.755	0.769	0.425	0.455	-0.449	0.129	1.557	0.212	0.002
F18C	0.546	0.599	0.390	-0.450	-0.231	0.015	26.631	0.000	0.020
F20C	0.477	0.714	0.294	0.209	0.471	0.525	11.798	0.001	0.011
F21C	0.796	0.653	0.497	0.526	0.032	0.192	8.349	0.004	0.009
F25C	0.785	0.473	0.542	0.496	0.448	0.095	44.780	0.000	0.015
F28C	0.689	0.492	0.489	0.294	0.556	0.160	11.558	0.001	0.012
F32C	0.425	0.680	0.285	-0.686	-0.814	0.084	2.101	0.147	0.003
F35C	0.700	0.753	0.403	0.367	-0.481	0.102	4.717	0.030	0.004
F37C	0.664	0.765	0.418	0.302	-0.133	0.375	1.325	0.250	0.004
F39C	0.857	0.484	0.635	0.882	0.292	0.052	6.080	0.014	0.004
F41C	0.488	0.824	0.285	-0.245	-1.425	0.037	10.959	0.001	0.006
F44C	0.607	0.726	0.376	-0.027	-0.717	0.050	2.989	0.084	0.003
F01	0.797	0.782	0.532	0.671	-0.023	0.249	4.059	0.044	0.008
F05	0.391	0.609	0.412	0.133	0.739	0.351	8.940	0.003	0.015
F06	0.377	0.388	0.249	-1.120	1.459	0.035	32.082	0.000	0.035
F08	0.599	0.543	0.509	-0.006	0.655	0.152	2.295	0.130	0.009
F09	0.697	0.507	0.591	0.437	0.738	0.128	18.076	0.000	0.021
F10	0.476	0.294	0.460	0.462	1.456	0.112	3.117	0.077	0.009
F13	0.233	0.757	0.224	-0.076	1.139	0.617	4.996	0.025	0.012
F14	0.683	0.622	0.488	0.497	0.771	0.337	6.650	0.010	0.016
F16	0.789	0.635	0.624	0.914	0.507	0.208	2.499	0.114	0.005
F17	0.561	0.245	0.544	0.538	1.364	0.047	15.722	0.000	0.016
F19	0.465	0.705	0.311	-0.704	-0.602	0.058	37.721	0.000	0.036
F22	0.691	0.502	0.648	1.036	0.776	0.161	7.274	0.007	0.012
F23	0.498	0.568	0.387	-0.462	0.488	0.103	13.273	0.000	0.022
F24	0.440	0.878	0.278	-0.083	-0.232	0.613	5.678	0.017	0.007
F26	0.537	0.417	0.468	0.374	1.357	0.190	13.586	0.000	0.027
F27	0.803	0.611	0.617	0.384	0.424	0.078	1.920	0.166	0.008
F29	0.453	0.610	0.377	-0.092	0.722	0.321	3.851	0.050	0.010
F30	0.502	0.427	0.522	0.641	1.104	0.162	17.064	0.000	0.021
F31	0.669	0.673	0.488	-0.141	0.178	0.099	13.360	0.000	0.022
F33	0.862	0.681	0.672	0.788	0.394	0.133	1.887	0.170	0.005
F34	0.747	0.415	0.603	0.526	1.033	0.078	12.001	0.001	0.019

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
F36	0.736	0.651	0.583	0.665	0.636	0.218	2.669	0.102	0.010
F38	0.711	0.622	0.527	0.117	0.530	0.128	5.459	0.019	0.014
F40	0.502	0.594	0.489	0.295	0.704	0.258	2.417	0.120	0.009
F42	0.603	0.715	0.395	-0.372	-0.101	0.128	8.196	0.004	0.016
F43	0.527	0.483	0.449	-0.154	0.967	0.134	6.200	0.013	0.016

Table 193: Factor Analysis Communalities, CTT Difficulty (*p*) and Discrimination (*r*), IRT Log of Discrimination (*log a*), Difficulty *b*, Guessing *c*, Goodness of Fit Chi-Square and Probability, and Root Mean Square Error (*RMSE*) for the Science - Chemistry Test

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
Q01C	0.220	0.705	0.323	-0.138	-0.293	0.302	5.782	0.016	0.005
Q03C	0.496	0.703	0.427	0.344	-0.105	0.188	4.504	0.034	0.004
Q06C	0.701	0.357	0.358	-0.222	0.983	0.117	0.605	0.437	0.002
Q08C	0.638	0.416	0.498	0.409	0.615	0.134	7.695	0.006	0.006
Q09C	0.324	0.661	0.392	0.167	0.119	0.301	2.242	0.134	0.004
Q11C	0.351	0.735	0.399	0.224	-0.390	0.224	3.094	0.079	0.004
Q14C	0.515	0.465	0.435	-0.078	0.276	0.002	105.827	0.000	0.025
Q15C	0.515	0.731	0.428	0.394	-0.217	0.164	0.882	0.348	0.002
Q17C	0.361	0.501	0.541	0.551	0.465	0.161	2.291	0.130	0.005
Q19C	0.736	0.363	0.584	0.506	0.624	0.051	9.779	0.002	0.007
Q23C	0.556	0.829	0.291	-0.139	-1.286	0.087	11.039	0.001	0.005
Q25C	0.499	0.679	0.446	0.117	-0.474	0.057	2.918	0.088	0.004
Q27C	0.591	0.408	0.445	0.201	0.939	0.128	12.665	0.000	0.013
Q28C	0.549	0.676	0.521	0.603	-0.208	0.080	2.521	0.112	0.004
Q33C	0.615	0.541	0.420	0.197	0.426	0.178	1.238	0.266	0.003
Q34C	0.521	0.699	0.444	0.396	-0.251	0.207	0.806	0.369	0.001
Q35C	0.530	0.729	0.419	0.168	-0.448	0.035	44.885	0.000	0.013
Q38C	0.570	0.609	0.403	0.154	0.041	0.179	2.638	0.104	0.004
Q02	0.420	0.557	0.587	0.271	0.601	0.089	8.142	0.004	0.018
Q04	0.689	0.440	0.366	-0.171	1.153	0.207	3.054	0.081	0.011
Q05	0.672	0.463	0.509	0.153	0.781	0.159	6.876	0.009	0.014
Q07	0.448	0.276	0.409	0.116	1.516	0.126	7.456	0.006	0.016
Q10	0.692	0.338	0.267	-0.880	1.752	0.076	9.891	0.002	0.018
Q12	0.275	0.546	0.642	0.396	0.281	0.020	8.731	0.003	0.010
Q13	0.491	0.821	0.403	-0.040	-0.780	0.070	19.458	0.000	0.015
Q16	0.477	0.647	0.533	0.083	0.028	0.074	2.653	0.103	0.008
Q18	0.610	0.777	0.441	0.113	-0.213	0.277	4.010	0.045	0.009
Q20	0.400	0.503	0.509	0.038	0.826	0.140	7.814	0.005	0.018
Q21	0.619	0.608	0.516	0.042	0.142	0.085	9.951	0.002	0.016
Q22	0.140	0.811	0.409	0.358	-0.158	0.426	3.095	0.079	0.007
Q24	0.658	0.657	0.550	0.292	0.445	0.226	2.418	0.120	0.009
Q26	0.701	0.510	0.330	-0.424	1.052	0.218	2.694	0.101	0.010
Q29	0.685	0.445	0.315	-0.788	1.061	0.101	5.155	0.023	0.015
Q30	0.664	0.679	0.587	0.148	0.092	0.083	9.692	0.002	0.020
Q31	0.587	0.470	0.453	-0.052	0.992	0.172	4.635	0.031	0.013
Q32	0.402	0.557	0.500	0.041	0.709	0.137	4.640	0.031	0.014
Q36	0.402	0.831	0.507	0.506	-0.188	0.300	6.107	0.013	0.014
Q37	0.320	0.817	0.336	-0.244	-0.727	0.287	10.353	0.001	0.009
Q39	0.538	0.653	0.555	0.343	0.228	0.147	2.647	0.104	0.007

Item	Communality	*p*	*r*	log_*a*	*b*	*c*	Chi	Prob	RMSE
Q40	0.534	0.588	0.599	0.242	0.651	0.138	8.790	0.003	0.025
Q41	0.369	0.634	0.600	0.159	0.191	0.071	3.404	0.065	0.008
Q42	0.425	0.380	0.483	0.029	1.126	0.089	3.377	0.066	0.010
Q43	0.454	0.344	0.486	0.144	1.219	0.086	10.958	0.001	0.017
Q44	0.473	0.457	0.501	0.378	1.053	0.185	3.656	0.056	0.014

Appendix E. Source of Challenge

This appendix deals with assessment items that the content area specialists felt to be problematic, either because they could be interpreted as being biased, because their wording was ambiguous, or because they assessed areas of knowledge that were outside the Chilean Ministry of Education's high school content standards. The results are presented in a tabular form and are organized by test. The left-hand column indicates the assessment item number, and the right-hand column presents the content area specialist's rationale for flagging that problem.

Table 194: Sources of Challenge in Content by PSU Test

PSU Test: Language and Communication	
Item #	**Explanation**
3	This item shows a potential bias against female test takers. It is a question about whether people in communicative situations are symmetrical. In it, all of the potentially "powerful" people are portrayed as male or described using words that leave their gender unknown, but none are specifically female. This could be resolved by making two of the examples women.
8	This item creates a potential bias against Hindu test takers. In this example fragment, the fantastical world created involves Brahma, a Hindu god, in an unofficial "re-created" creation story by Gustavo Adolfo Becquer. Hindu test-takers might take offense at the reinterpretation of their deity. This could be resolved by using a piece that does not involve an actual religion.
29	There is more than one way to organize the ideas for this item.
31	This item shows a potential bias against female test takers. The focus is on the appearance of a woman as an example of something to identify an "idealized" tone. This could be resolved by using a passage that idealizes an object or abstract notion instead of a woman.
38	This item shows a potential bias against female test takers. The passage is all about someone's annoyance about "that fat (female) maid, who talked, and talked, and talked, and talked." It may be seen to perpetuate the stereotype of women as maids whose opinions are unimportant, especially when a woman is not physically attractive (etc). We suggest selecting a passage that does not rely on gender stereotypes.
64-67	The passage is an interview with Ridley Scott, in which he discusses a violent movie (Gladiator), a movie about a controversial war (Black Hawk Down), and September 11th. We suggest a more emotionally neutral passage.
68	This item shows a potential bias against female test takers. The writing in this sample discusses social security and labor policy, but highlights women as a particular beneficiary / impoverished group. This could be resolved by having a selection based on a policy debate in which women are not singled out.

PSU Test: Science – Biology	
Item #	**Explanation**
28	The test item involves interpreting a graph about tobacco consumption among teenagers and is perhaps not appropriate given the intended audience of the PSU. We suggest another topic for graphical interpretation.
PSU Test: Science – Chemistry	
Item #	**Explanation**
27	The question asks test-takers to identify the IUPAC-approved name of a chemical compound generated from a chemical reaction. While chemical reactions are addressed by the standards, the IUPAC naming system is not.

Appendix F. Interview Protocols

Pauta Entrevista: **Profesores de secundaria**

Instrucciones: Familiarícese con los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) del marco curricular de la enseñanza media para esta asignatura, así como el marco curricular para la formación diferenciada, los cuales se encuentran ubicados en su cuaderno de trabajo. Mientras lea la documentación, por favor considere y responda a las siguientes preguntas:

1. Utilizando los OF y los CMO, haga una marca de verificación o checkmark (√) para identificar esos objetivos que en su opinión están cubiertos en las aulas de su área temática.

2. Dibuje un círculo alrededor de esos OF y CMO en los cuales usted se sienta que los estudiantes tienen las mayor dificultades. Ofrezca sus pensamientos de por qué esto podría ser el caso.

3. Escriba un resumen en alto nivel del consenso grupal.

4. Por favor proporcione otros comentarios y/o experiencias que usted desee compartir.

Pauta Entrevista: **Profesores de universidad**

Instrucciones: Familiarícese con los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) del marco curricular de la enseñanza media para esta asignatura, así como el marco curricular para la formación diferenciada, los cuales se encuentran ubicados en su cuaderno de trabajo. Mientras lea la documentación, por favor considere y responda a las siguientes preguntas:

1. Utilizando los OF y los CMO, haga una marca de verificación o checkmark (√) para identificar esos objetivos que en su opinión más fuertemente definen las características del conocimiento que deben poseer al principio de su enseñanza universitaria.

2. Dibuje un círculo alrededor de esos OF y CMO en los cuales usted se sienta que los estudiantes tienen las mayor dificultades. Ofrezca sus pensamientos de por qué esto podría ser el caso.

3. ¿Hasta qué punto siente usted que la PSU como instrumento de selección universitaria, indica el nivel de preparación de los estudiantes universitarios al comienzo de sus estudios?

4. Escriba un breve resumen del consenso grupal.

5. Por favor proporcione otros comentarios y/o experiencias que usted desee compartir.

Pauta Entrevista: **Expertos de currículo**

Instrucciones: Familiarícese con los Objetivos Fundamentales (OF) y los Contenidos Mínimos Obligatorios (CMO) del marco curricular de la enseñanza media para esta asignatura (por ejemplo, Matemáticas, Lengua, Historia y Ciencias Sociales, Ciencias-Biología, Física, Química), así como el marco curricular para la formación diferenciada de dicha asignatura, los cuales se encuentran ubicados en su cuaderno de trabajo. Mientras lea la documentación, por favor considere y responda a las siguientes preguntas:

1. La batería de pruebas que componen la Prueba de Selección Universitaria (PSU) fundamenta su construcción en los CMO del Marco Curricular de la Enseñanza Media. El sistema de selección actual supone que los estudiantes que reciben los puntajes más altos representan las mejores posibilidades de éxito en el cumplimiento de las tareas que les exige la educación superior. ¿Está usted de acuerdo con esta declaración, o no? Por favor explique su respuesta.

2. ¿Cuál es su comprensión de la relación entre el "Marco Curricular de la Enseñanza Media" y el nivel de conocimiento requerido para que los estudiantes universitarios de primer ingreso tengan éxito?

3. ¿Existe alguna coincidencia entre los CMO y los entendidos niveles de conocimiento de la materia temática que deben poseer los estudiantes universitarios de primer ingreso para tener éxito? Si es así, ¿en qué grado? Si no es así, por favor explique.

4. Discuta sus opiniones con otros miembros de su grupo y escriba un breve resumen del consenso para su grupo.

Appendix G. Factorial Analysis of Variance of PSU Subtest by Year, Gender, Type, Region and Curricular Branch

Table 195: Language—Factorial Analyses of Variance of PSU Subtest

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Language	Year	8	12605894	123.41	0.00	0.03
Language	Gender	1	3871627	303.21	0.00	0.01
Language	Year*Gender	8	4608700	45.12	0.00	0.02
Language	Error	1457312	18608147301			
Language	Year	8	11492447	113.02	0.00	0.02
Language	Region	2	91002649	3579.78	0.00	0.07
Language	Year*Region	16	5191775	25.53	0.00	0.02
Language	Error	1452877	18467006409			
Language	Year	8	11181831	125.63	0.00	0.03
Language	Type	3	2388530003	71561.12	0.00	0.39
Language	Year*Type	16	17132087	96.24	0.00	0.03
Language	Error	1443266	16057528632			
Language	Year	8	12599786	146.80	0.00	0.03
Language	Branch	8	2918497215	34003.72	0.00	0.43
Language	Year*Branch	56	63794850	106.18	0.00	0.06
Language	Error	1457243	15634169003			
Language	Year	8	9000385	105.9033	0.00	0.03
Language	SES Quintile	4	1088661599	25619.53	0.00	0.29
Language	Year*SES	32	64519413	189.7924	0.00	0.07
Language	Error	1247179	13249225642			

Table 196: Mathematics—Factorial Analyses of Variance of PSU Subtest

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Mathematics	Year	8	15750545	156.60	0.00	0.03
Mathematics	Gender	1	341006831	27123.63	0.00	0.14
Mathematics	Year*Gender	8	1748698	17.39	0.00	0.01
Mathematics	Error	1457312	18321786806			
Mathematics	Year	8	14262296	139.48	0.00	0.03
Mathematics	Region	2	26213229	1025.42	0.00	0.04
Mathematics	Year*Region	16	8059230	39.41	0.00	0.02
Mathematics	Error	1452877	18570156530			
Mathematics	Year	8	13884194	158.86	0.00	0.03
Mathematics	Type	3	2701392017	82422.95	0.00	0.41
Mathematics	Year*Type	16	31412922	179.71	0.00	0.04
Mathematics	Error	1443266	15767564183			
Mathematics	Year	8	15748406	188.78	0.00	0.03
Mathematics	Branch	8	3388625663	40620.82	0.00	0.47
Mathematics	Year*Branch	56	80192992	137.33	0.00	0.07
Mathematics	Error	1457243	15195565221			
Mathematics	Year	8	10751810	129.1164	0.00	0.03
Mathematics	SES Quintile	4	1065516725	25591.16	0.00	0.29
Mathematics	Year*SES	32	64196078	192.7295	0.00	0.07
Mathematics	Error	1247179	12981924217			

Table 197: History—Factorial Analyses of Variance of PSU Subtest

PSU Test	Source	DF	SS	F	p	Effect Size (f)
History	Year	8	5246573	54.24	0.00	0.02
History	Gender	1	160850782	13304.24	0.00	0.12
History	Year*Gender	8	724484	7.49	0.00	0.01
History	Error	965740	11675976480			
History	Year	8	4805144	49.31	0.00	0.02
History	Region	2	72698204	2983.89	0.00	0.08
History	Year*Region	16	3164106	16.23	0.00	0.02
History	Error	962800	11728639200			
History	Year	8	4619871	52.78	0.00	0.02
History	Type	3	1258642497	38346.44	0.00	0.35
History	Year*Type	16	15929672	91.00	0.00	0.04
History	Error	956265	10462473994			
History	Year	8	5244450	60.46	0.00	0.02
History	Branch	8	1310273450	15105.73	0.00	0.35
History	Year*Branch	56	56826459	93.59	0.00	0.07
History	Error	965677	10470370057			
History	Year	8	4338676	51.36	0.00	0.02
History	SES Quintile	4	542374148	12840.95	0.00	0.23
History	Year*SES	32	37237308	110.20	0.00	0.06
History	Error	838565	8854796041			

Table 198: Science—Factorial Analyses of Variance of PSU Subtest

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Science	Year	8	7894308	80.58	0.00	0.03
Science	Gender	1	140501930	11473.28	0.00	0.12
Science	Year*Gender	8	10890328	111.16	0.00	0.03
Science	Error	786740	9634426672			
Science	Year	8	7237842	73.47	0.00	0.03
Science	Region	2	87805400	3564.97	0.00	0.10
Science	Year*Region	16	5183867	26.31	0.00	0.02
Science	Error	784725	9663898005			
Science	Year	8	7092981	81.51	0.00	0.03
Science	Type	3	1206128084	36959.87	0.00	0.38
Science	Year*Type	16	13131934	75.45	0.00	0.04
Science	Error	779956	8484219234			
Science	Year	8	7891607	94.18	0.00	0.03
Science	Branch	8	1519199761	18131.14	0.00	0.43
Science	Year*Branch	56	27160767	46.31	0.00	0.06
Science	Error	786674	8239379915			
Science	Year	8	6414935	76.74	0.00	0.03
Science	SES Quintile	4	504546812	12072.06	0.00	0.27
Science	Year*SES	32	31993102	95.69	0.00	0.07
Science	Error	656432	6858828136			

Appendix H. Trend Analysis of PSU Scores by Subtest and Subpopulation

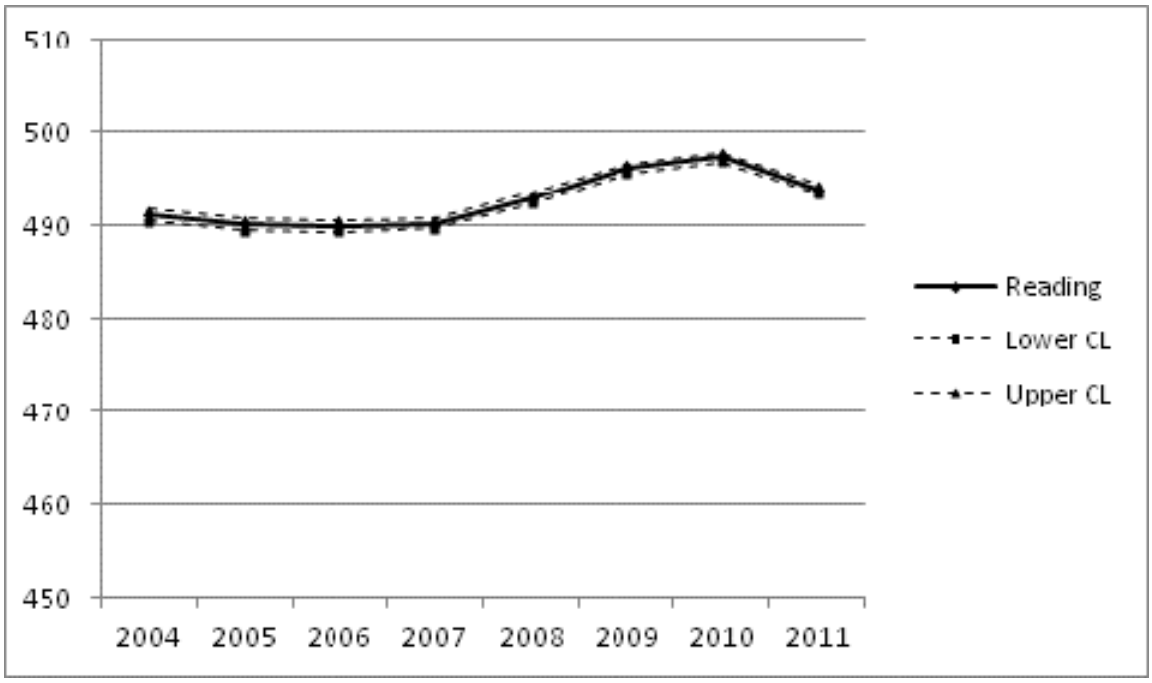


Figure 52: PSU Language and Communication Overall

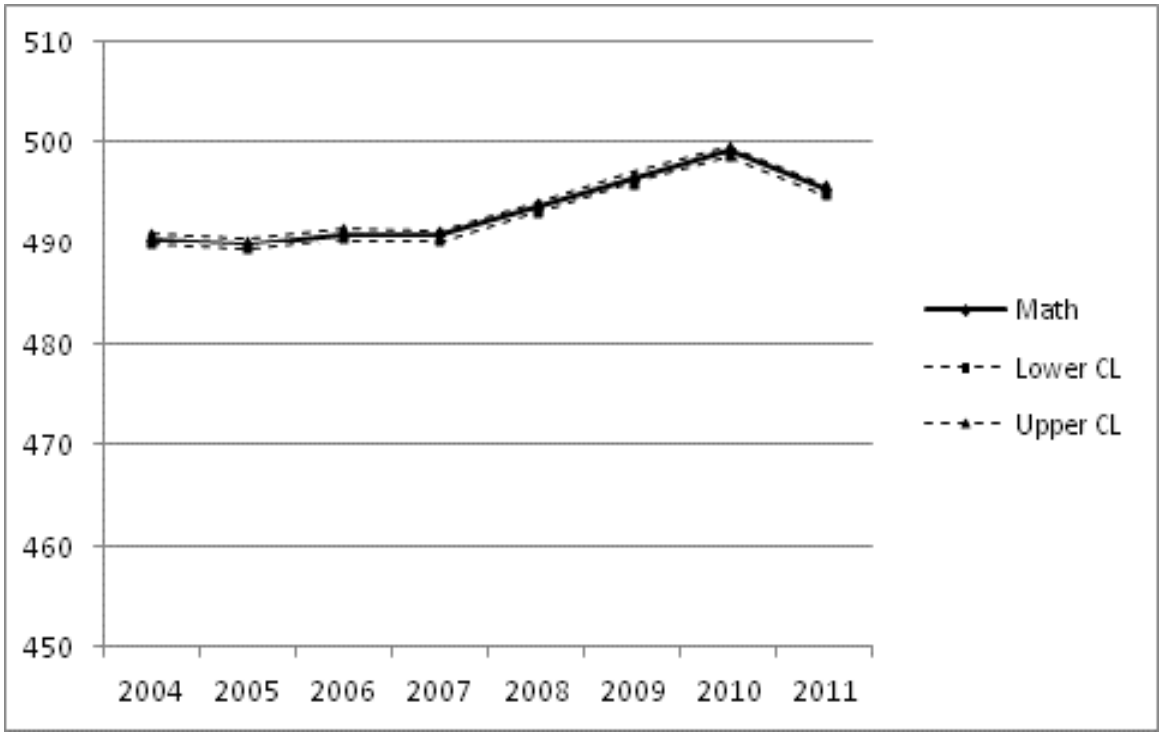


Figure 53: PSU Mathematics Overall

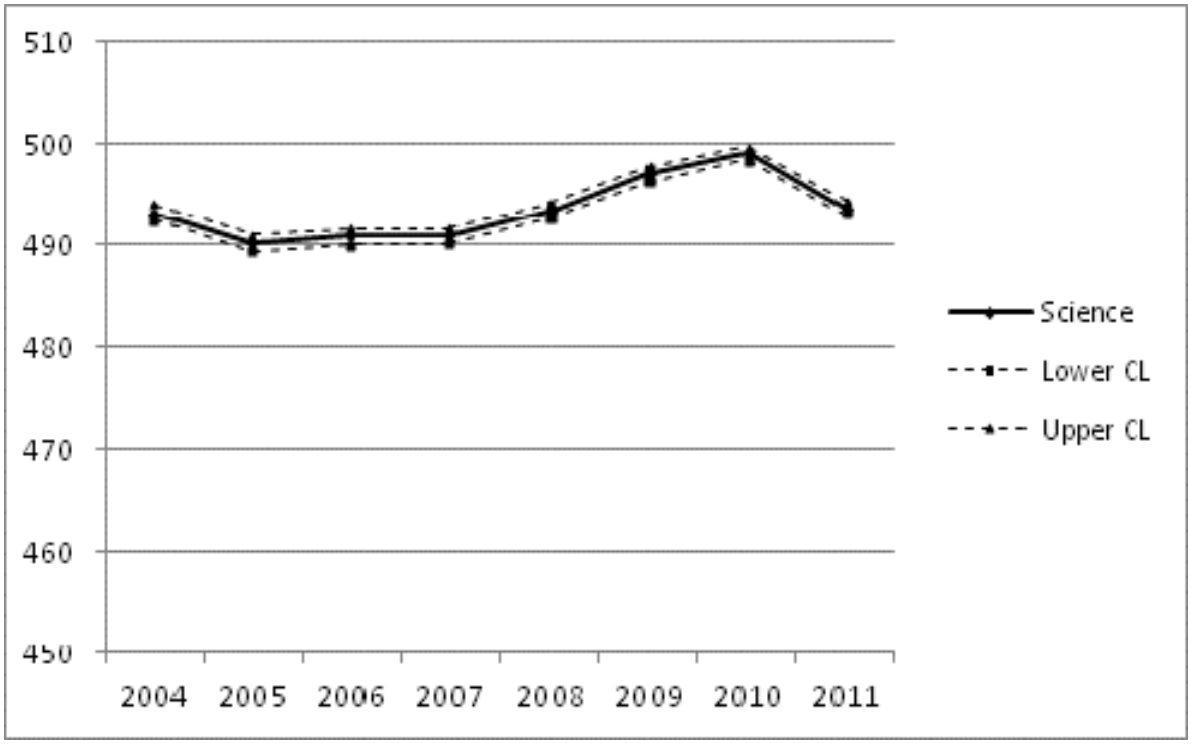


Figure 54: PSU Science Overall

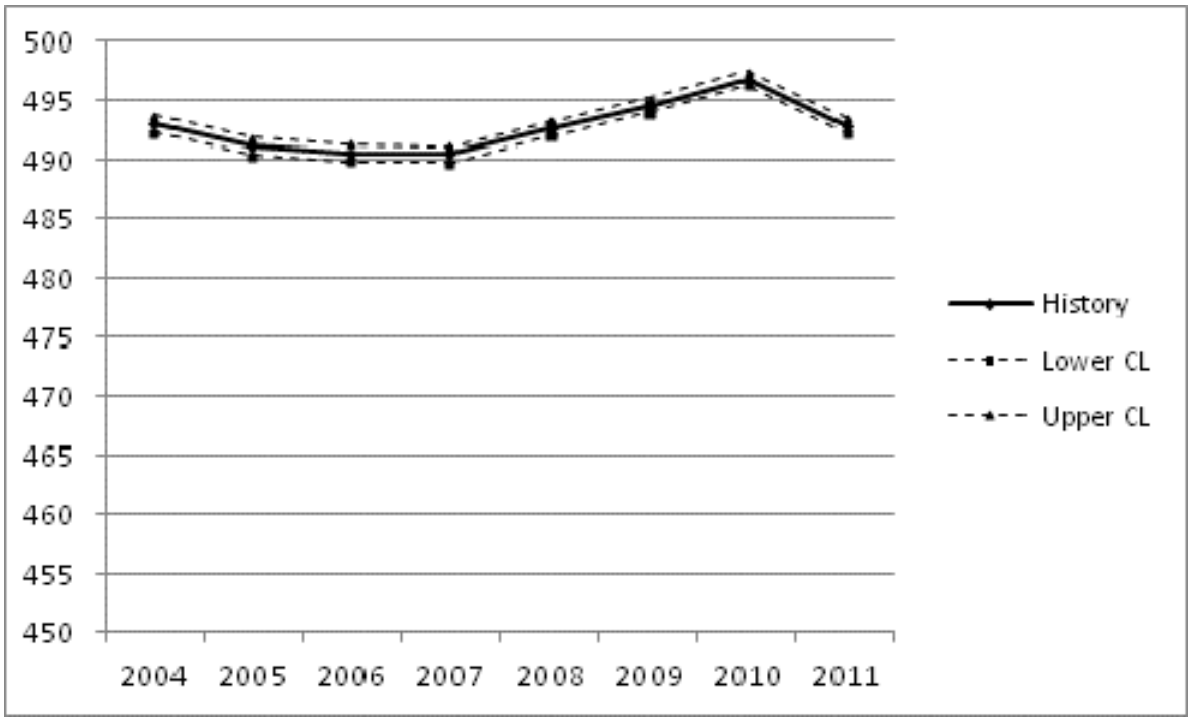


Figure 55: PSU History and Social Sciences Overall

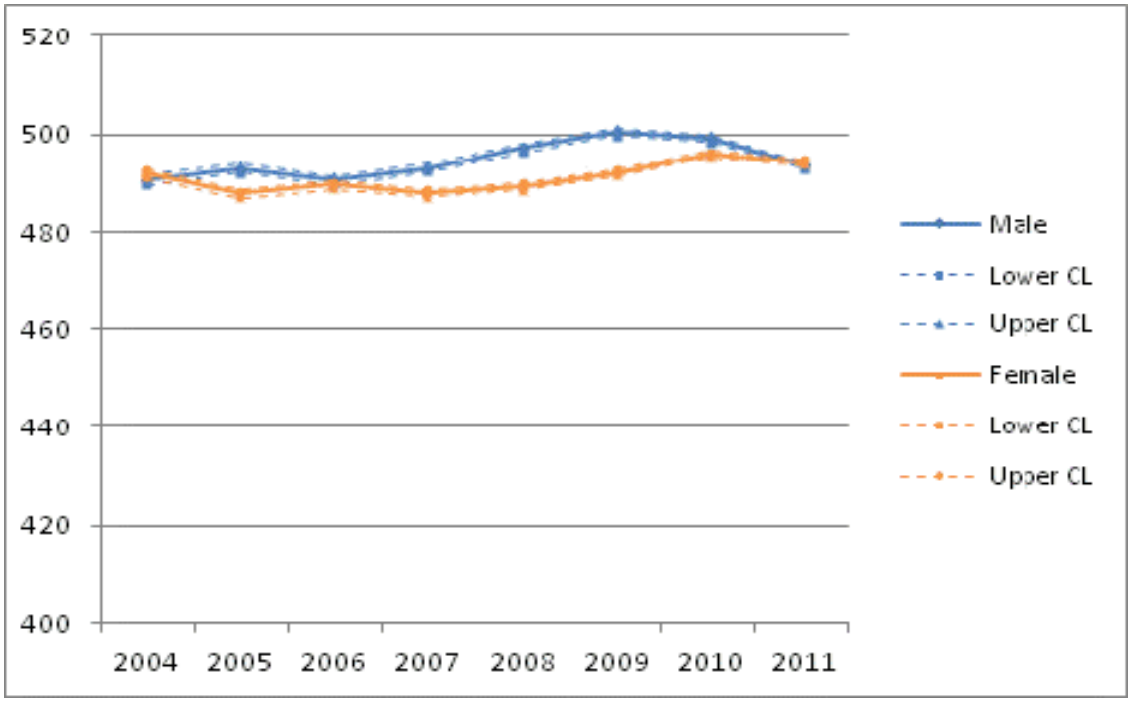


Figure 56: PSU Language and Communication by Gender

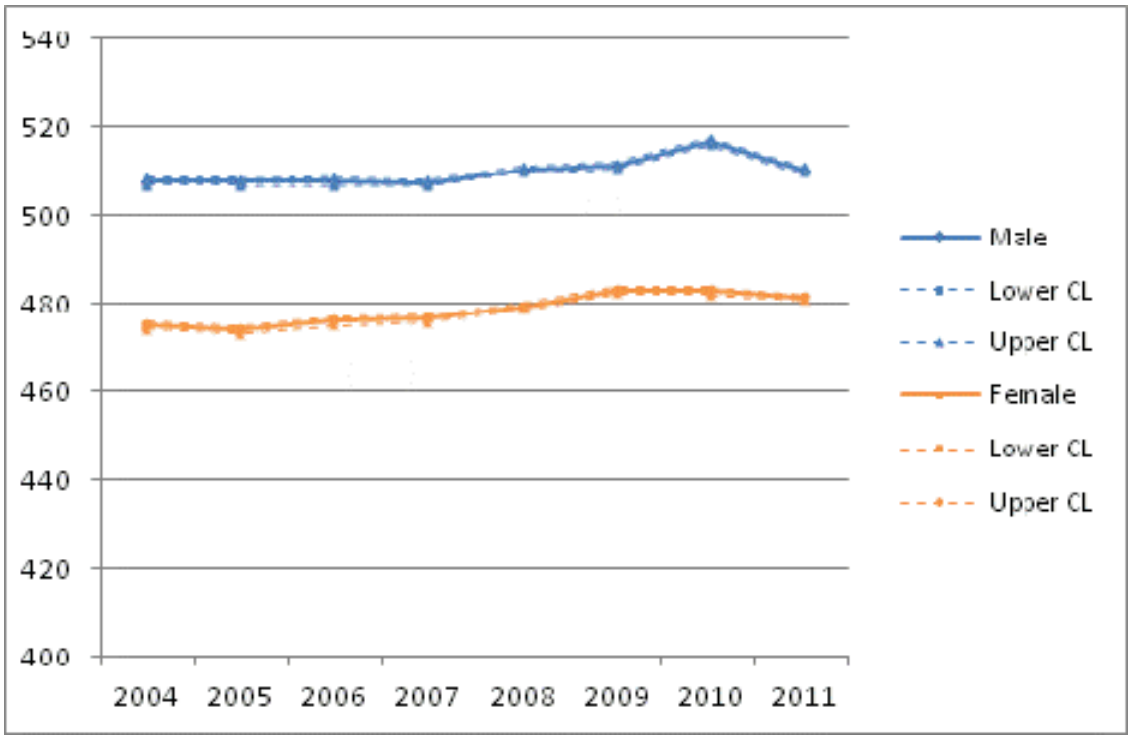


Figure 57: PSU Mathematics by Gender

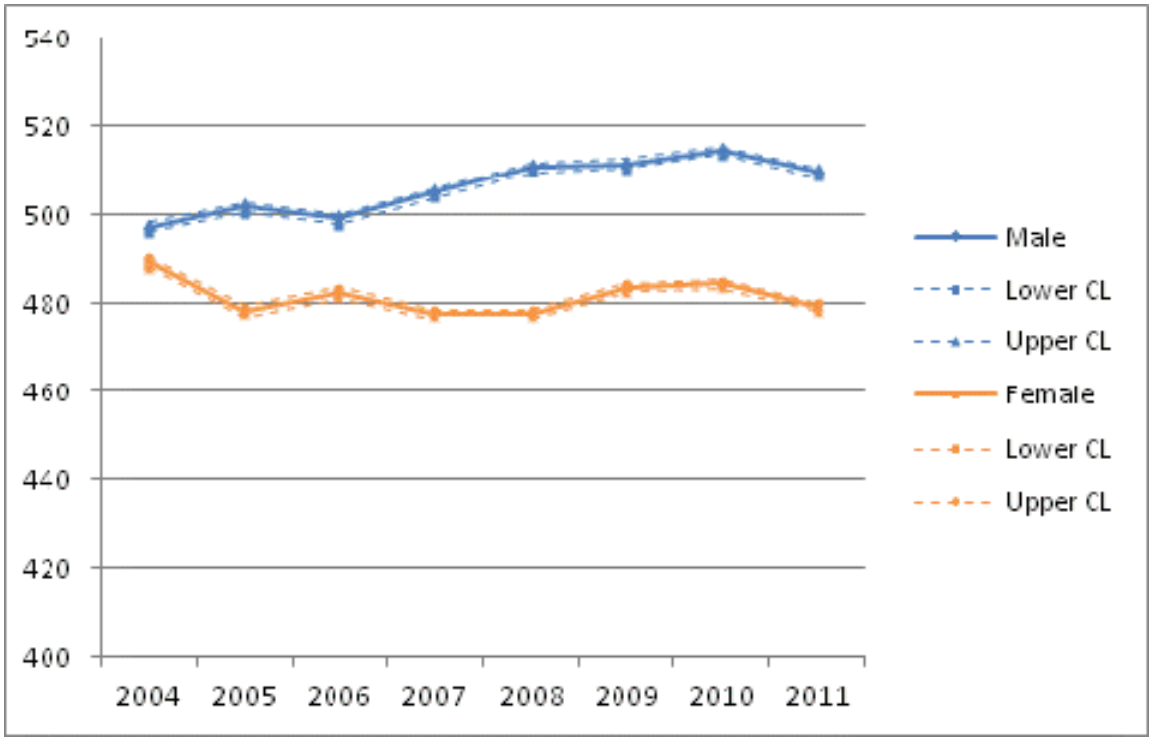


Figure 58: PSU Science by Gender

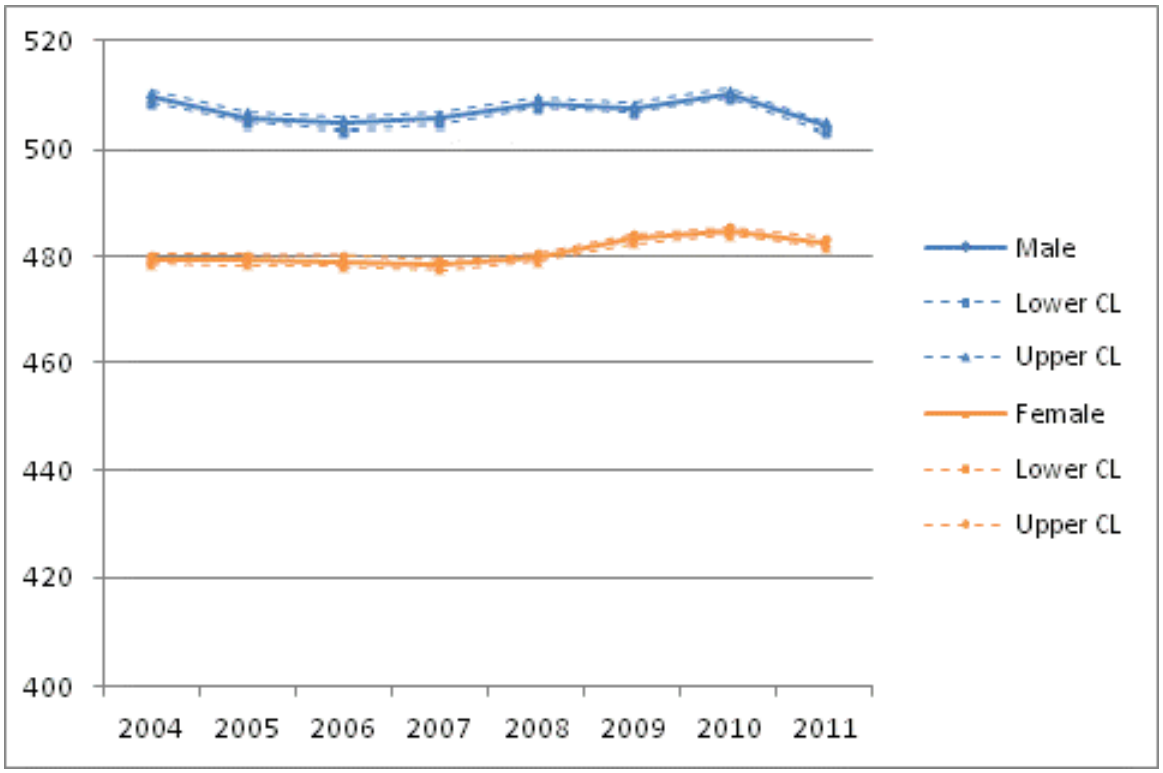


Figure 59: PSU History and Social Sciences by Gender

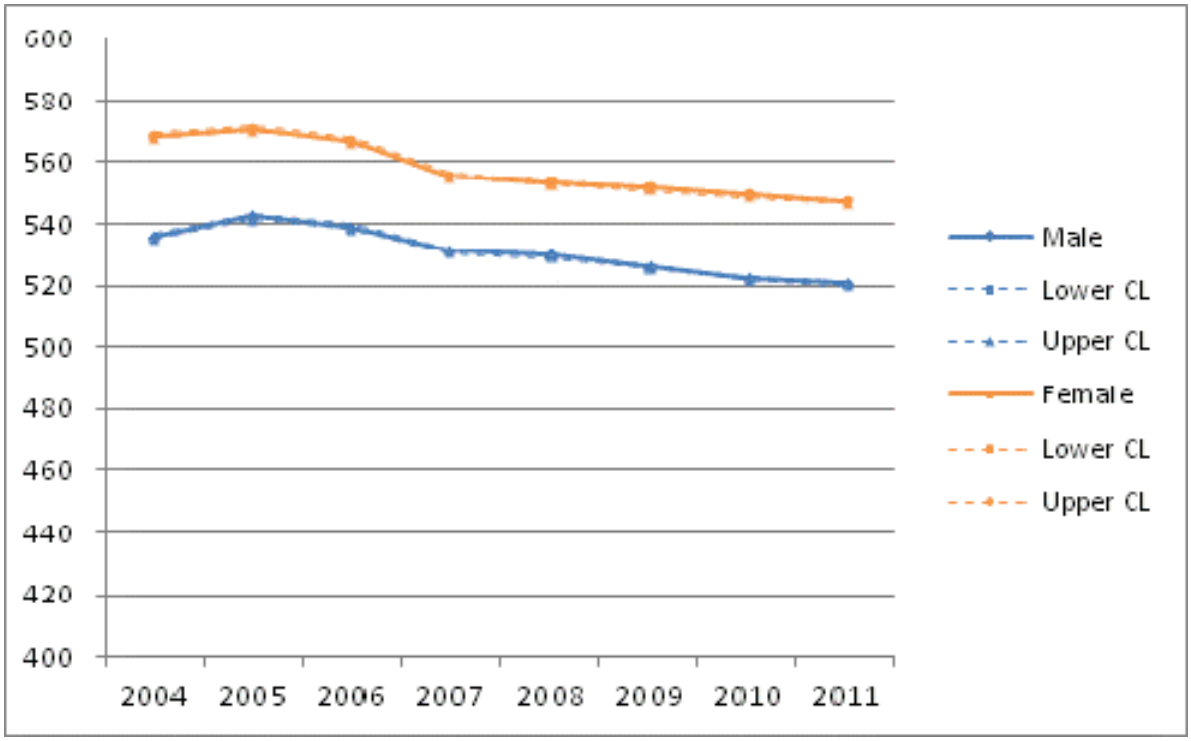


Figure 60: NEM by Gender

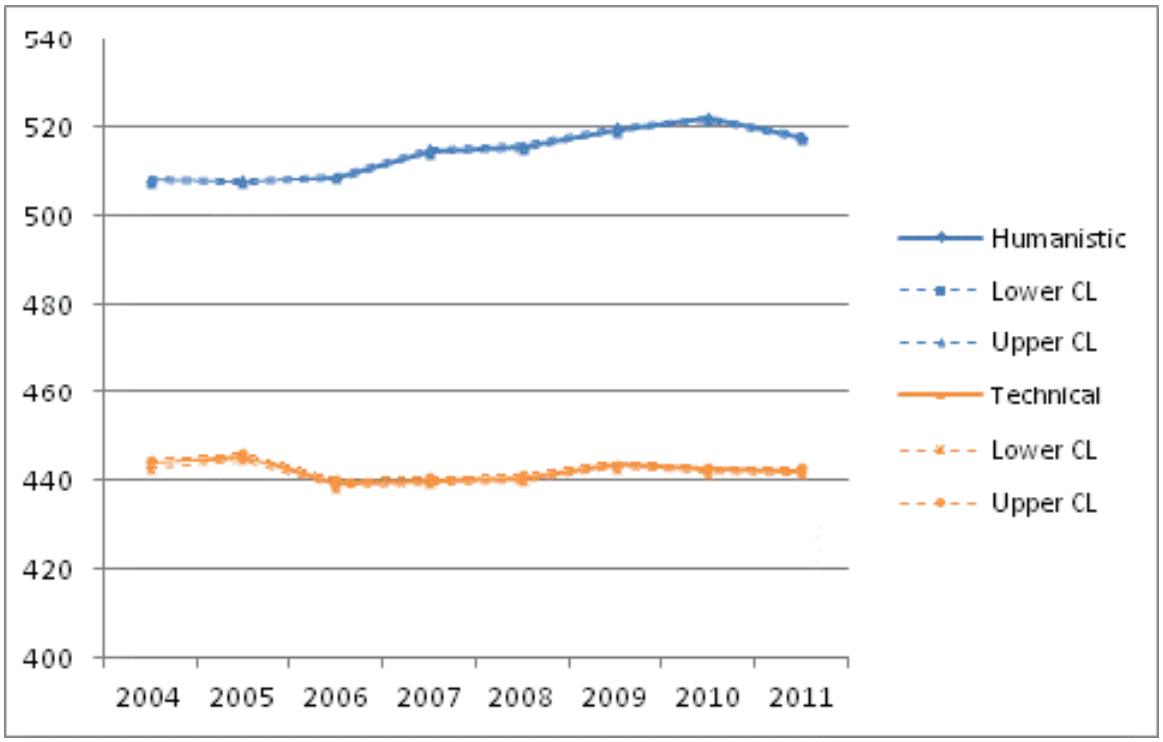


Figure 61: Language and Communication by Curricular Branch

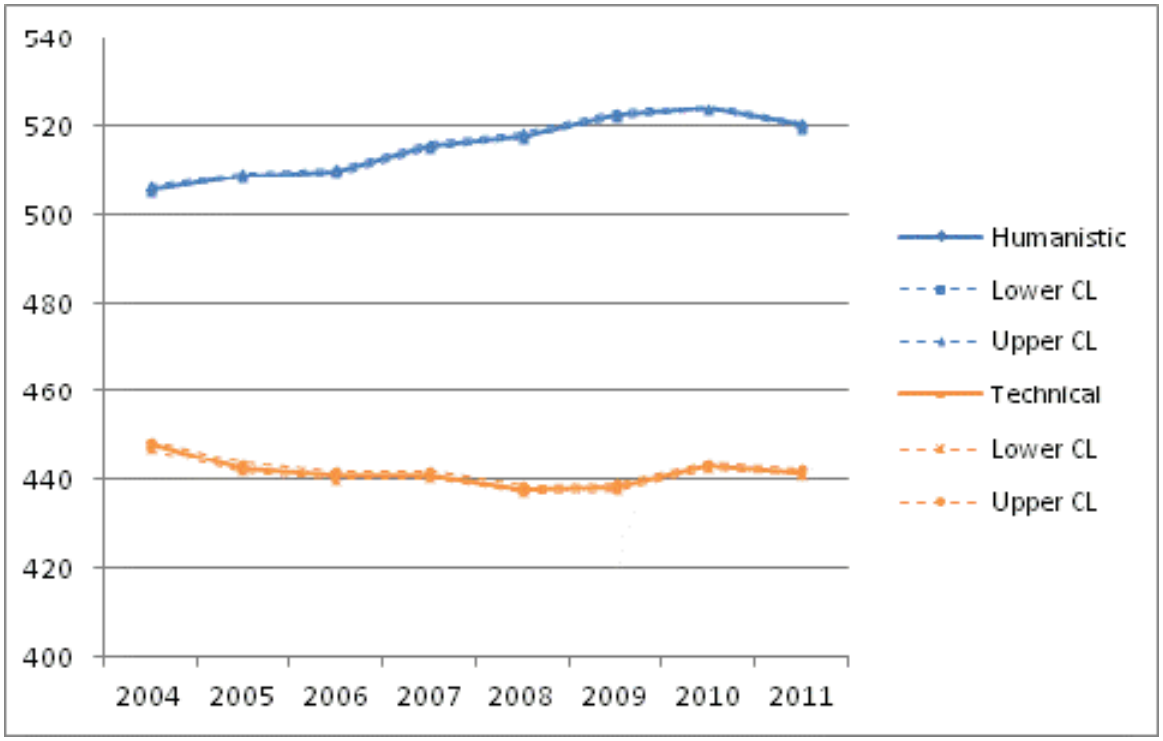


Figure 62: Mathematics by Curricular Branch

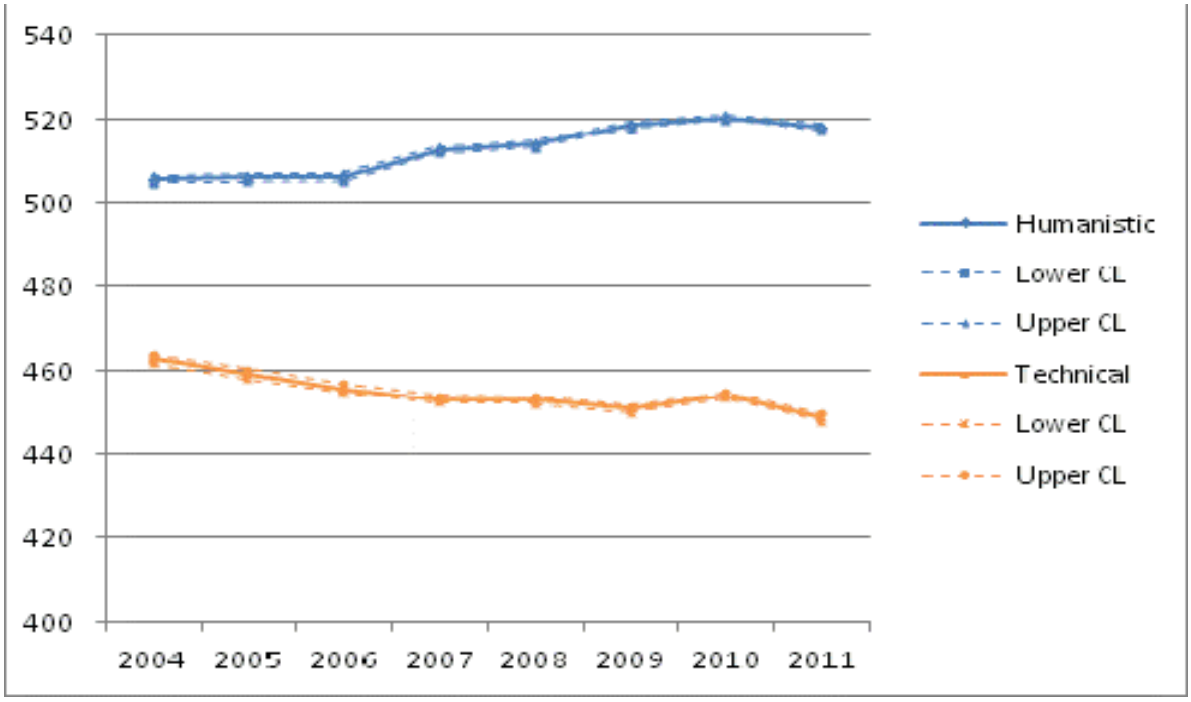


Figure 63: History and Social Sciences by Curricular Branch

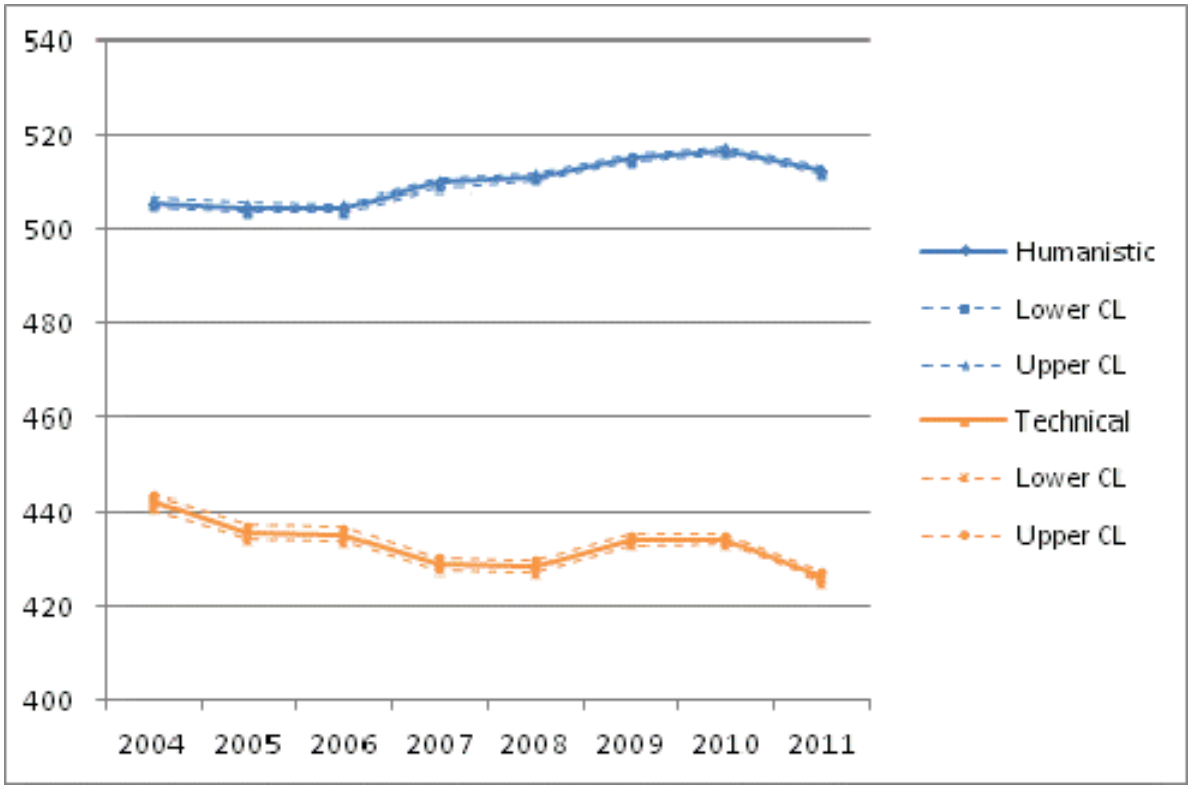


Figure 64: Science by Curricular Branch

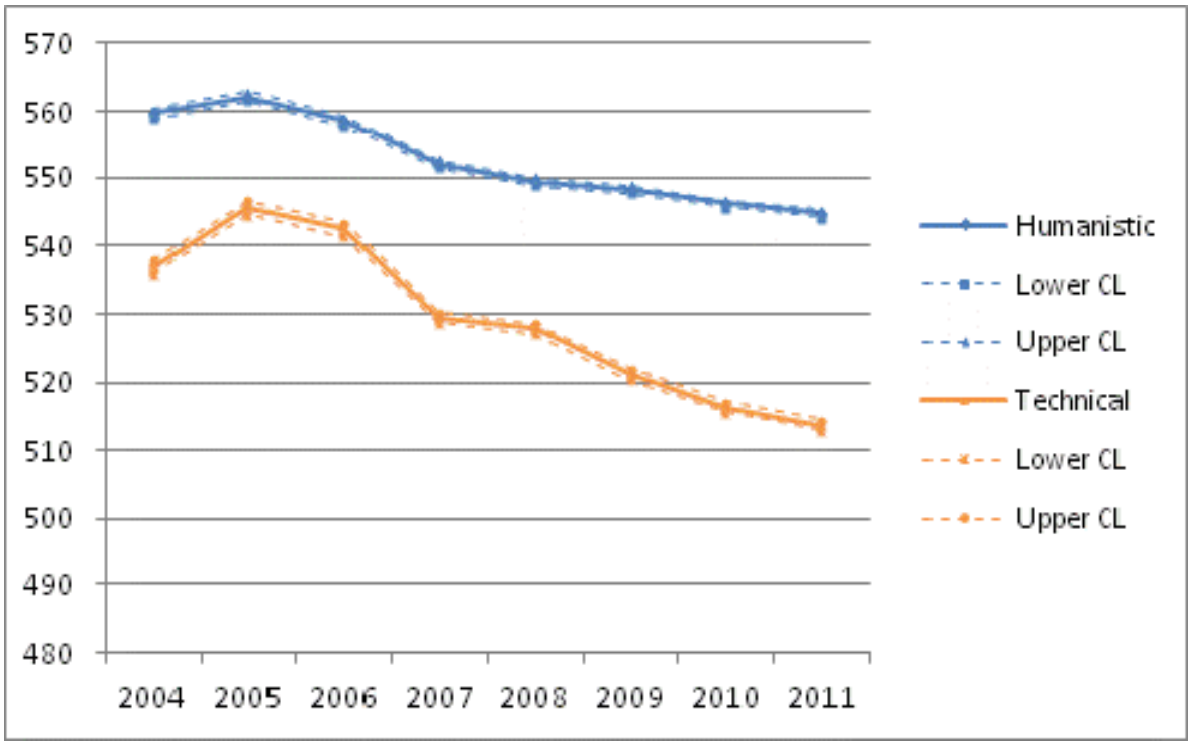


Figure 65: NEM by Curricular Branch

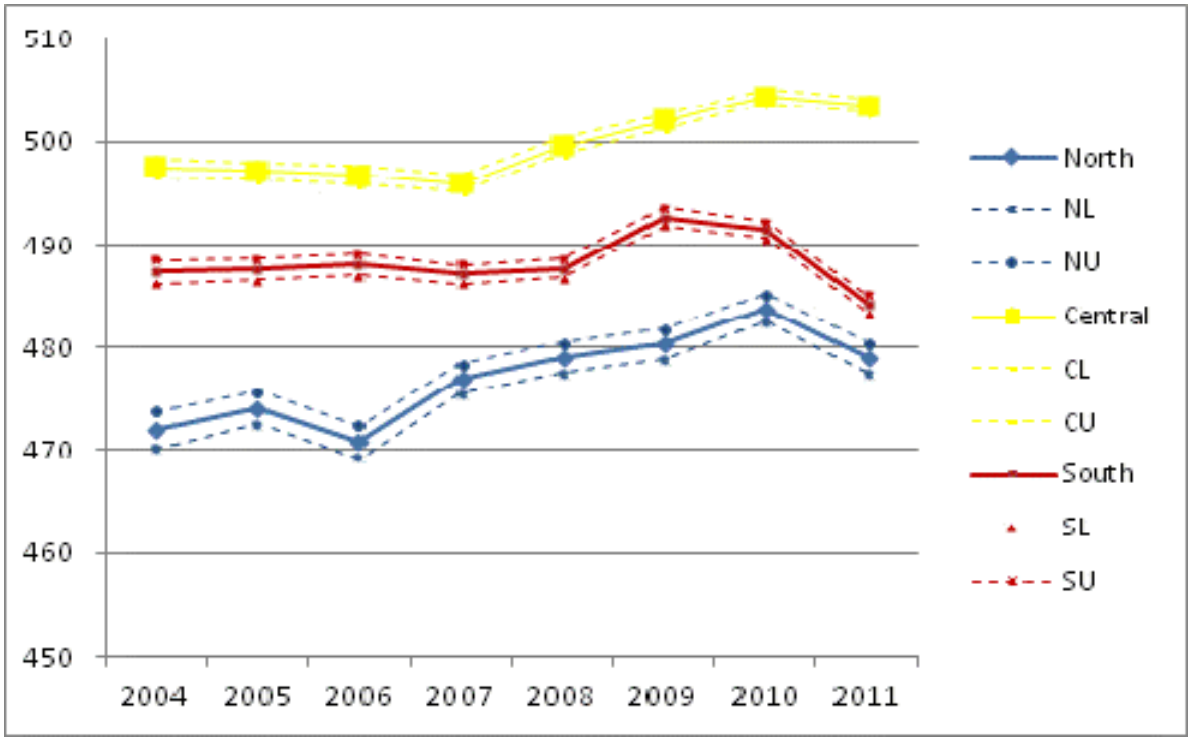


Figure 66: Language and Communication by Region

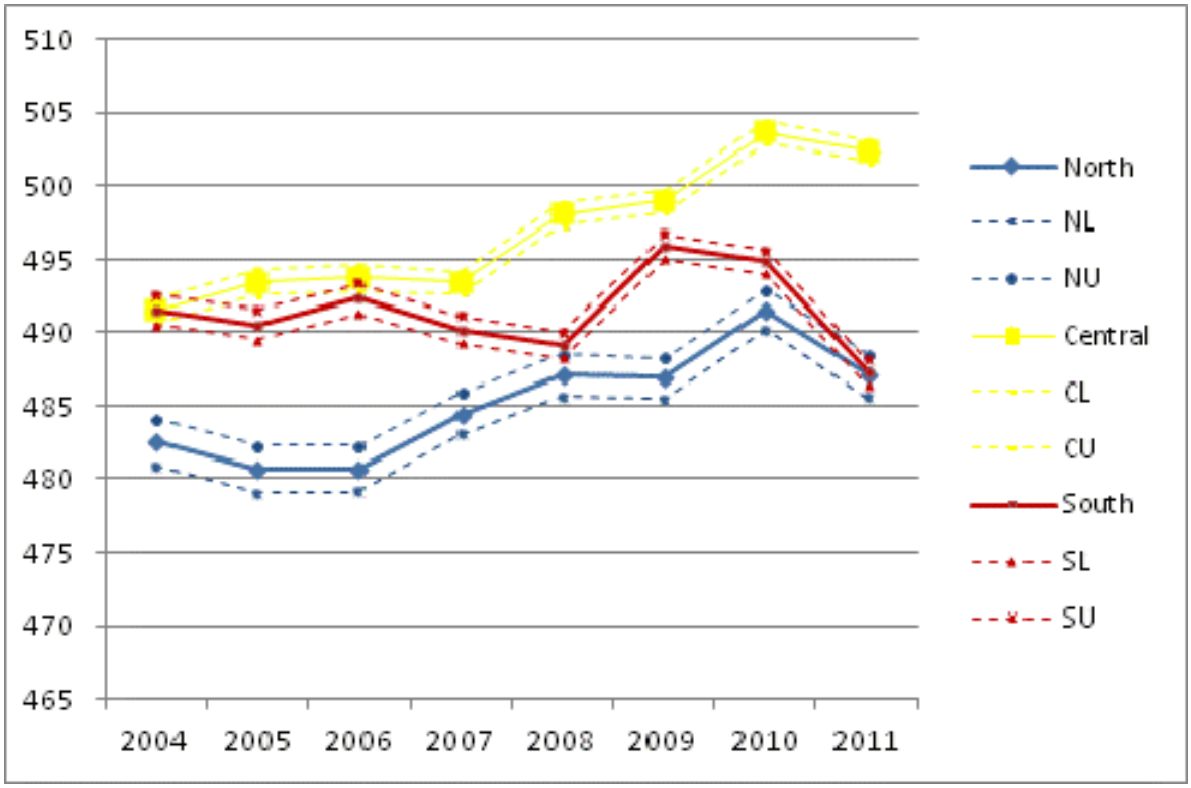


Figure 67: Mathematics by Region

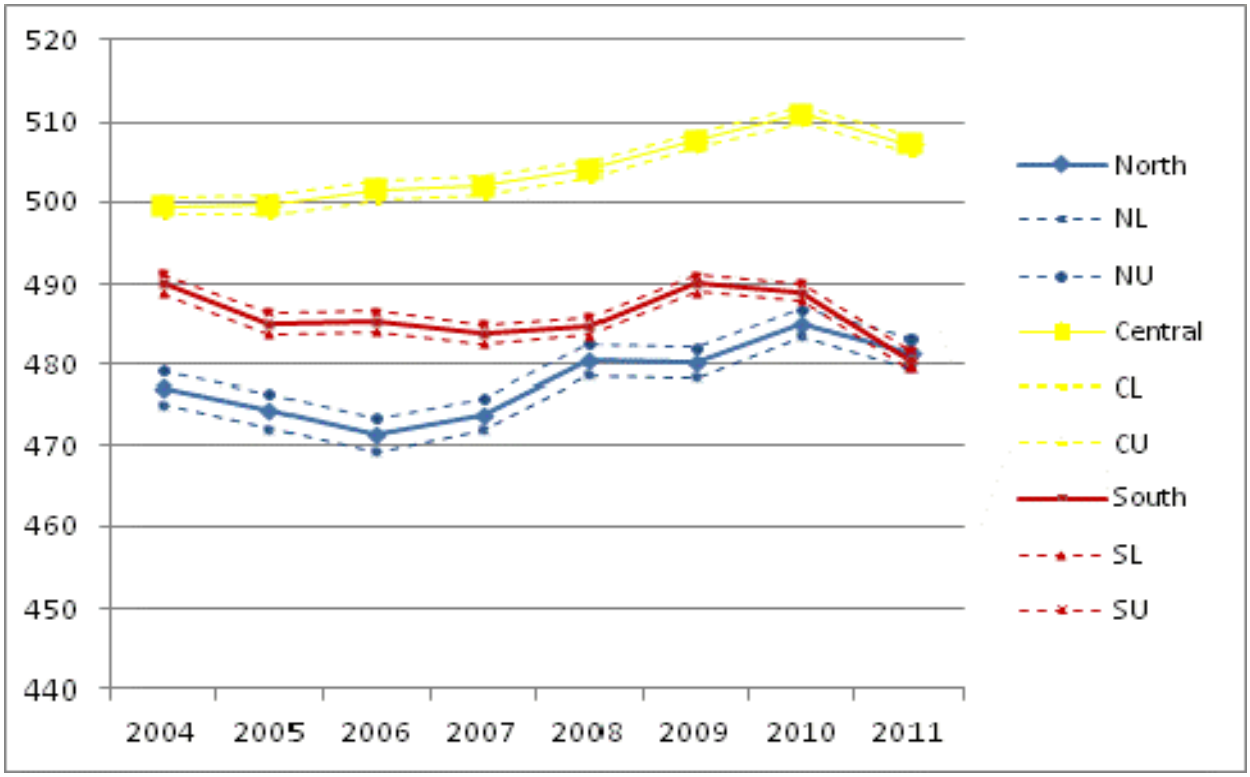


Figure 68: Science by Region

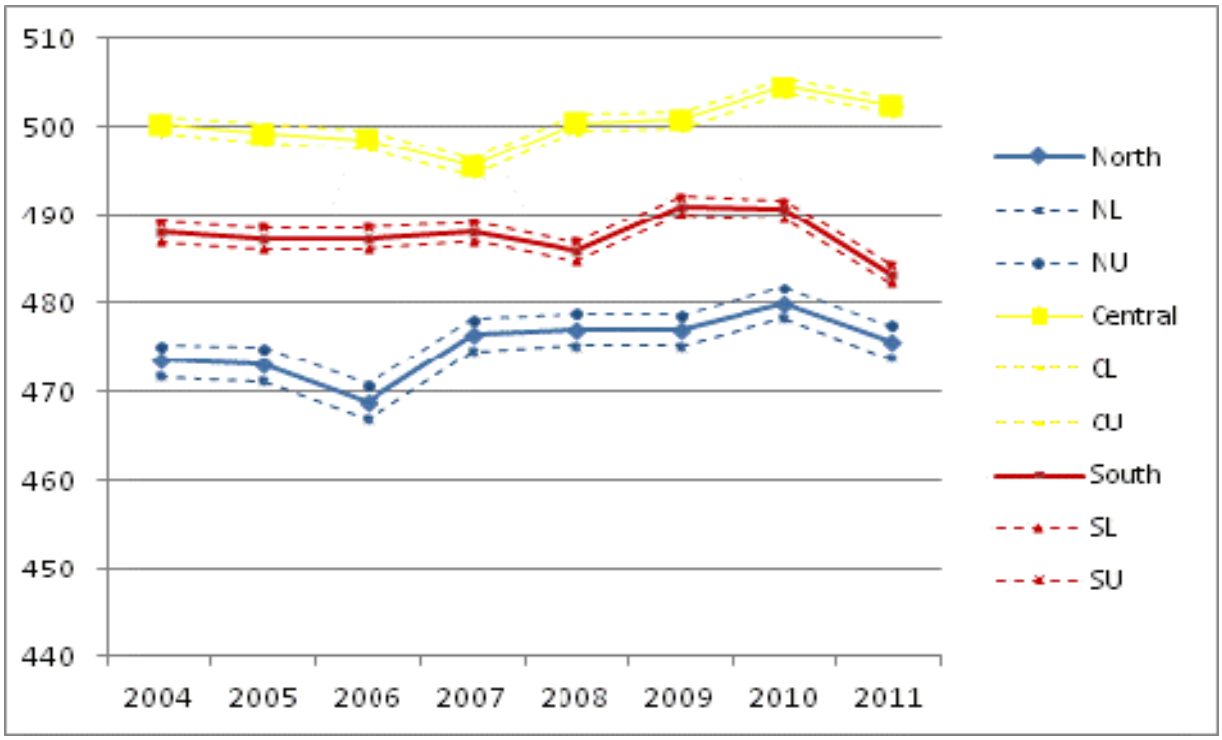


Figure 69: History and Social Sciences by Region

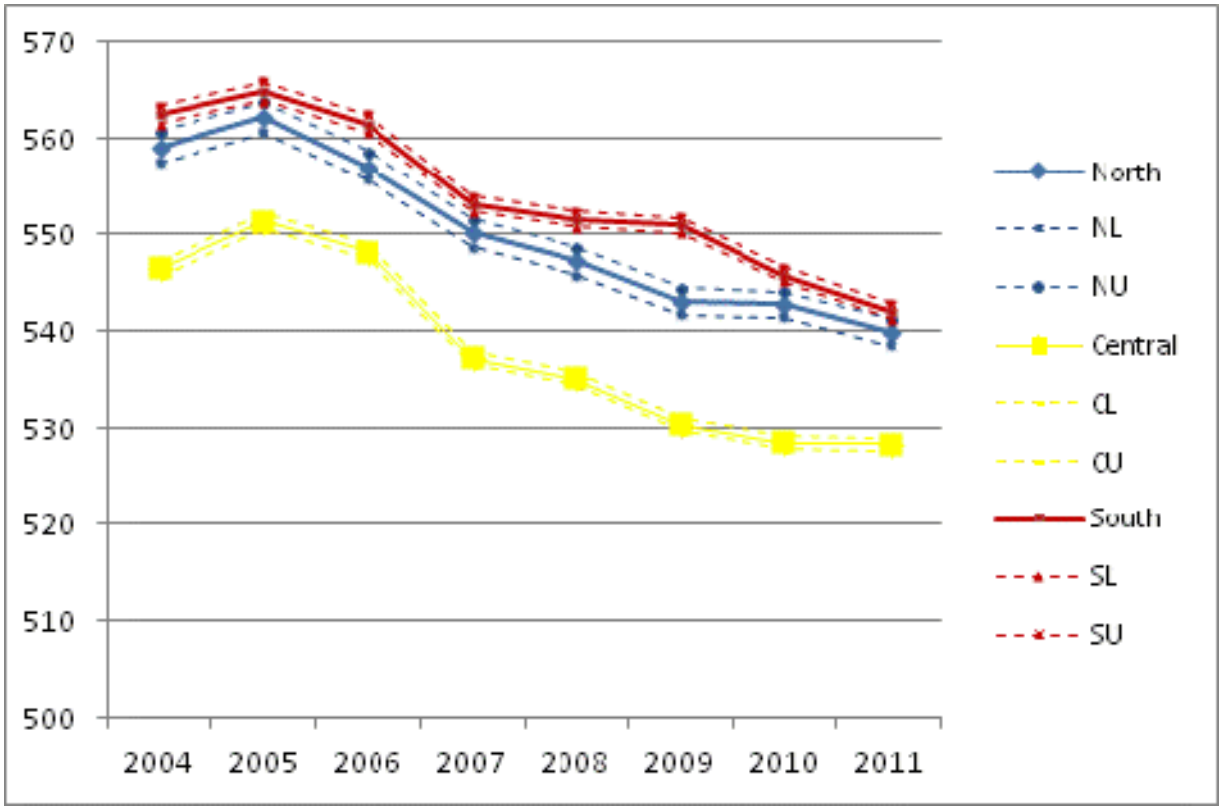


Figure 70: NEM by Region

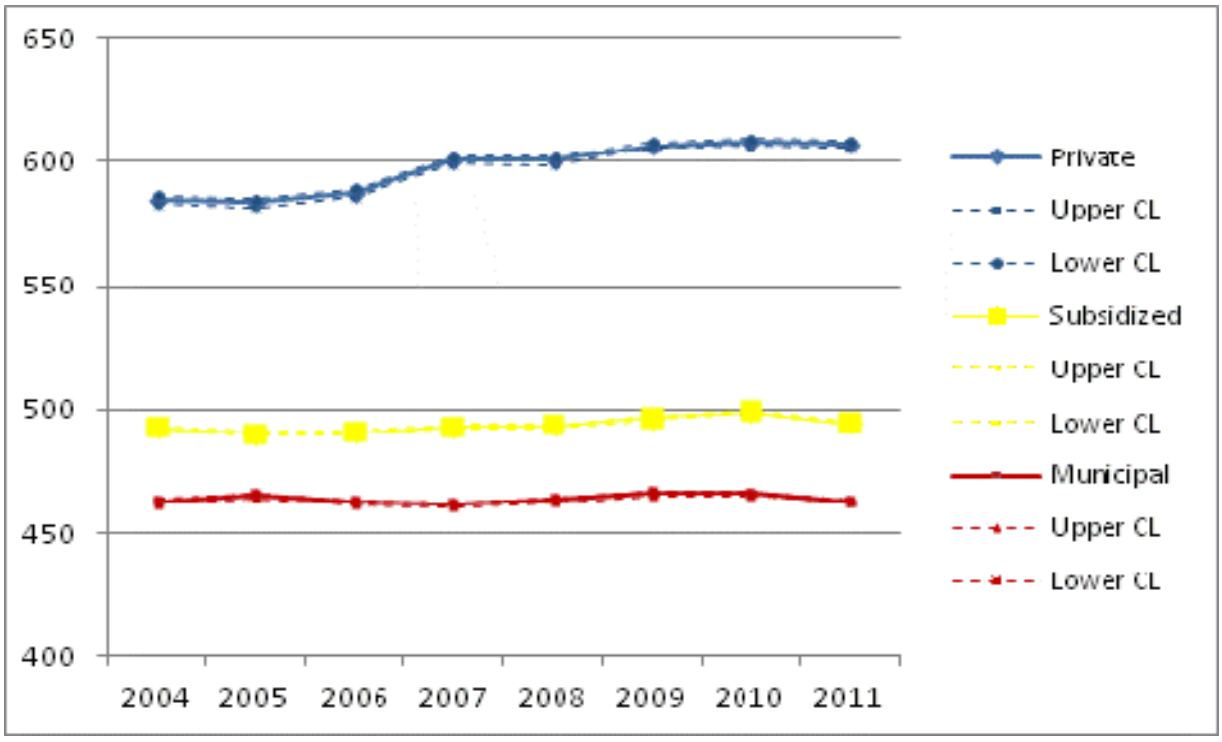


Figure 71: Language and Communication by School Funding

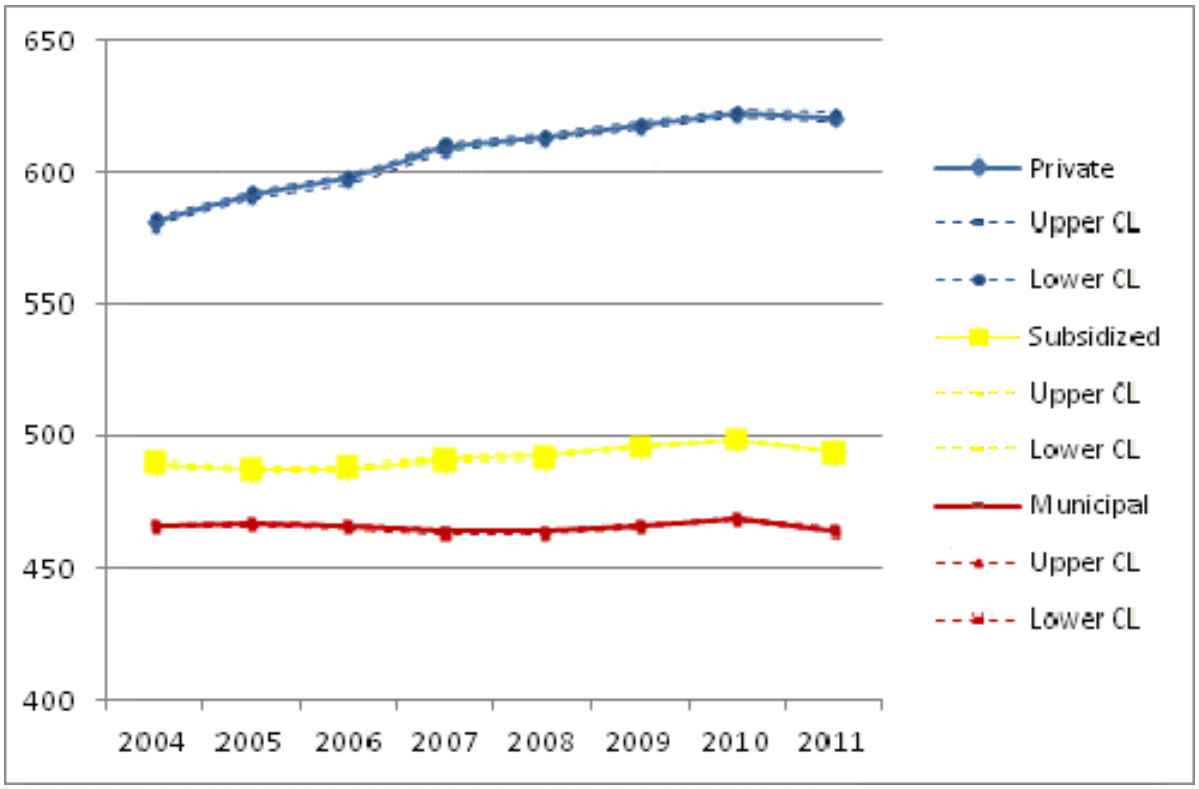


Figure 72: Mathematics by School Funding

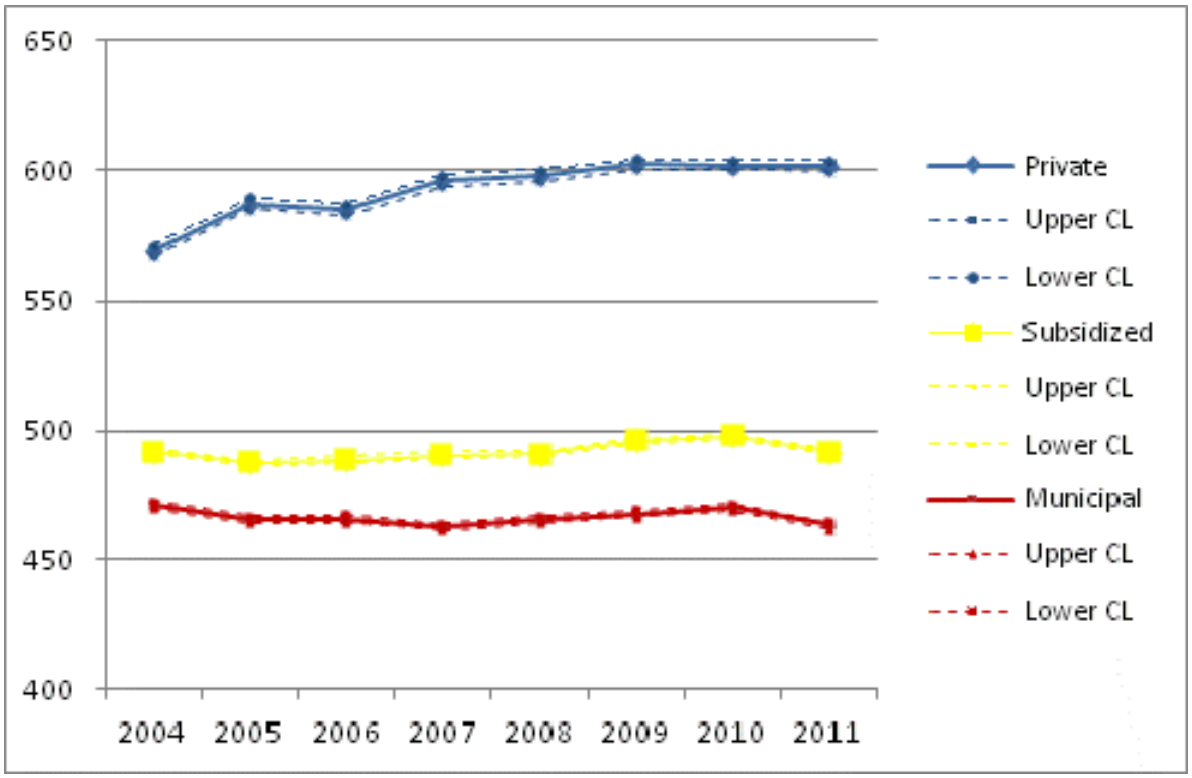


Figure 73: Science by School Funding

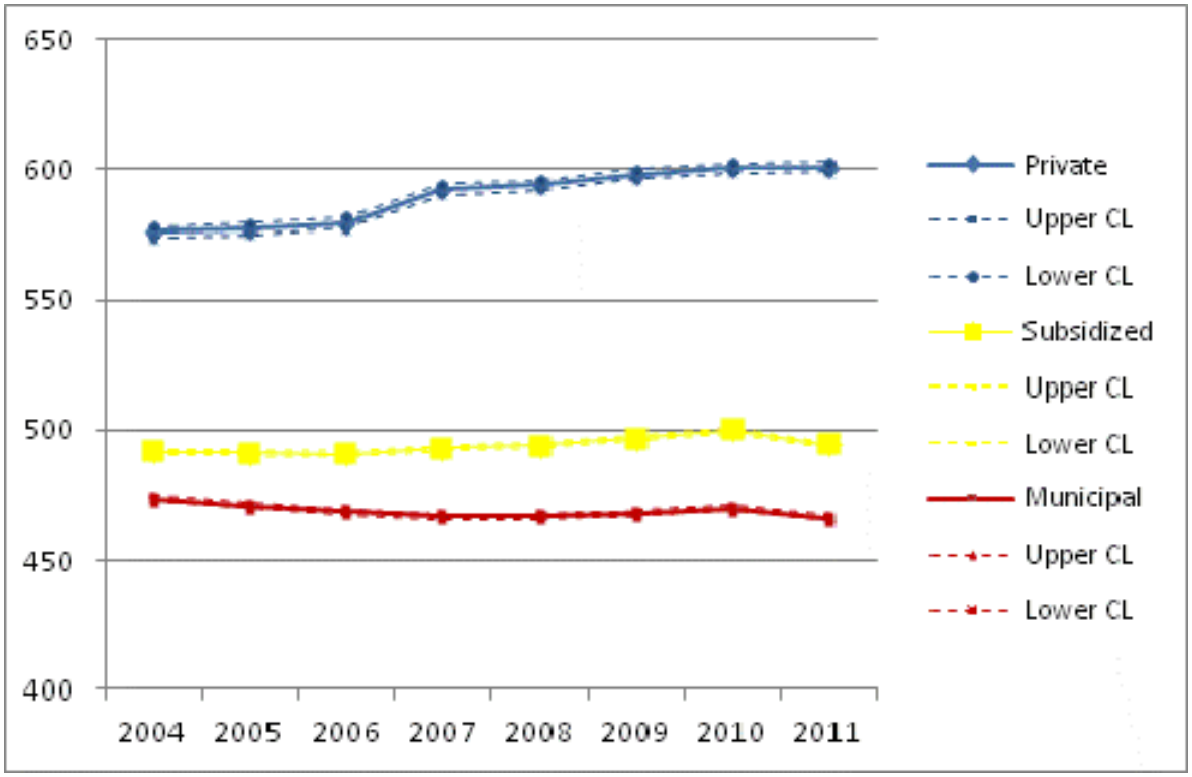


Figure 74: History and Social Sciences by School Funding

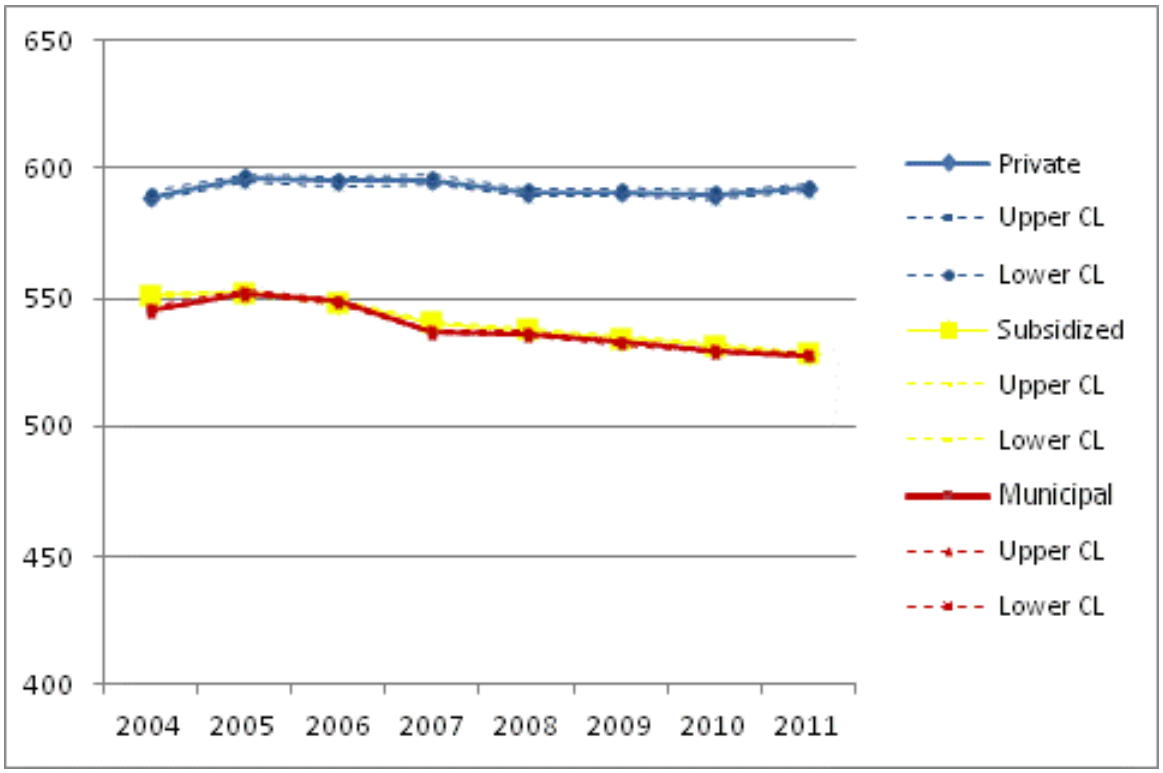


Figure 75: NEM by School Funding

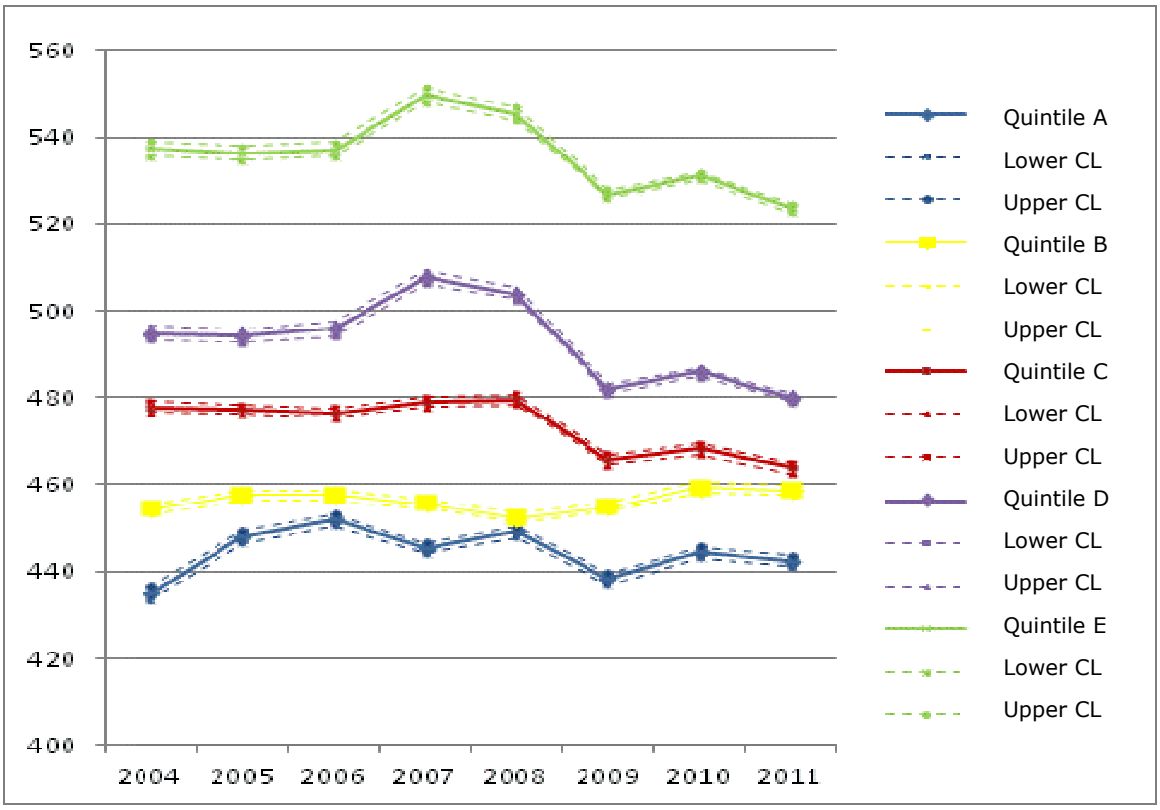


Figure 76: Language and Communication by Socioeconomic Status

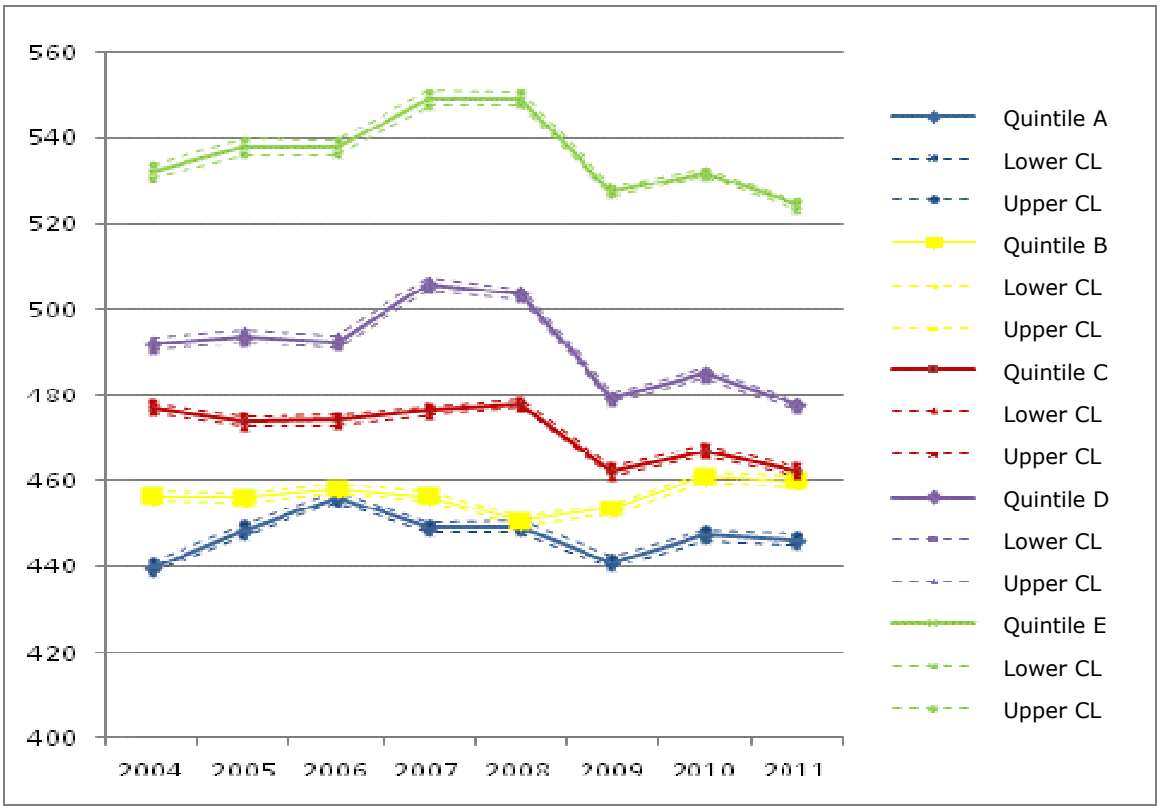


Figure 77: Mathematics by Socioeconomic Status

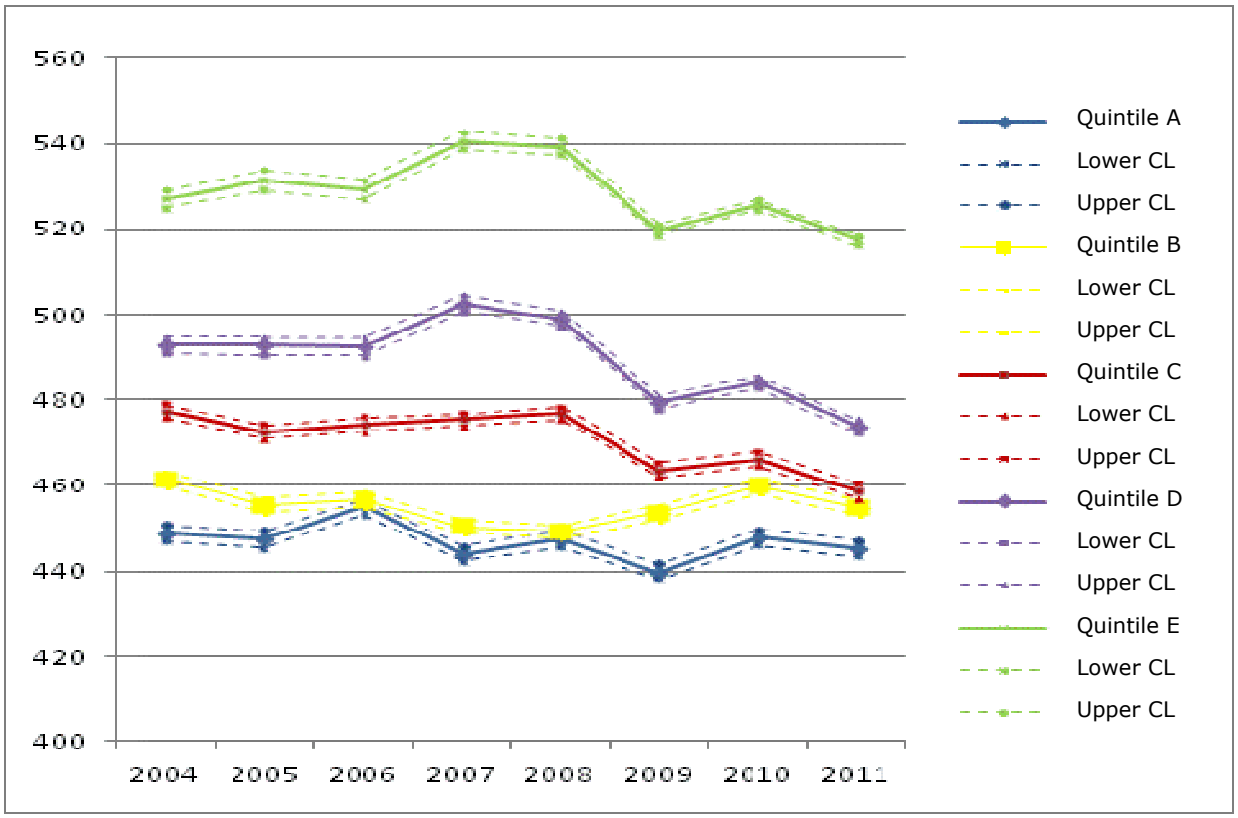


Figure 78: Science by Socioeconomic Status

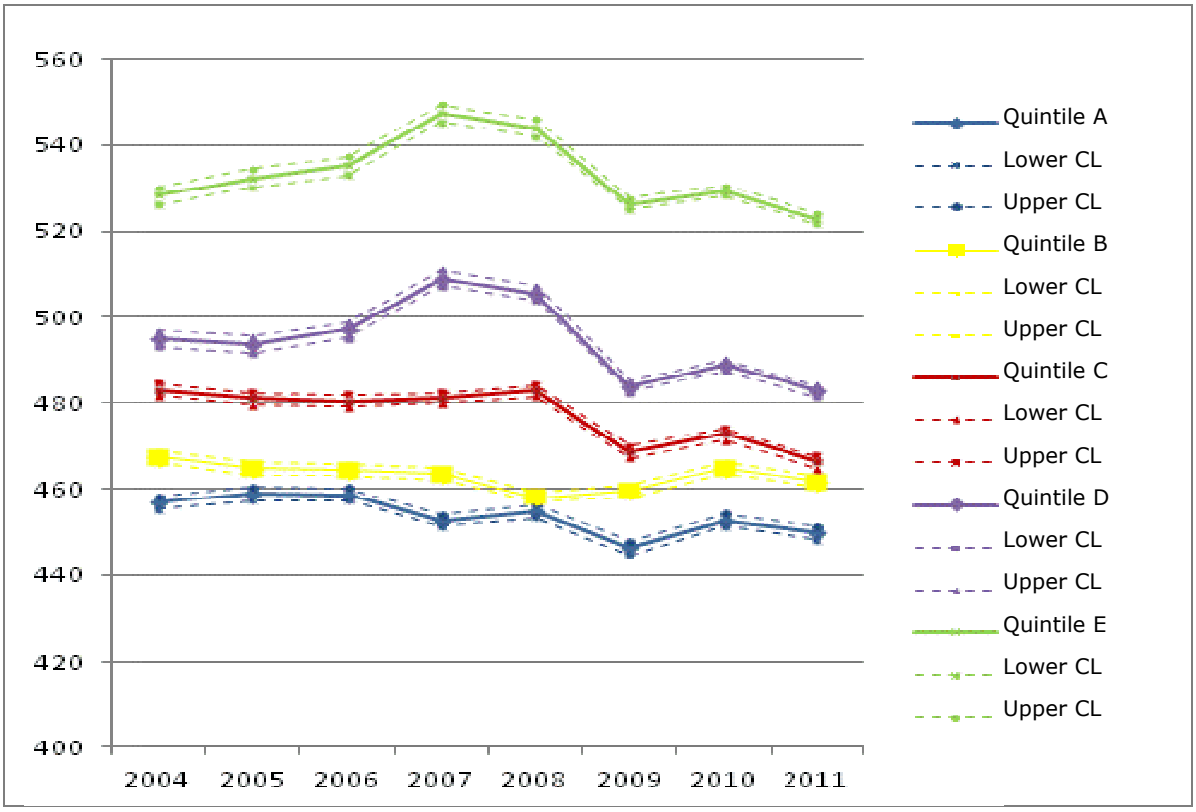


Figure 79: History and Social Sciences by Socioeconomic Status

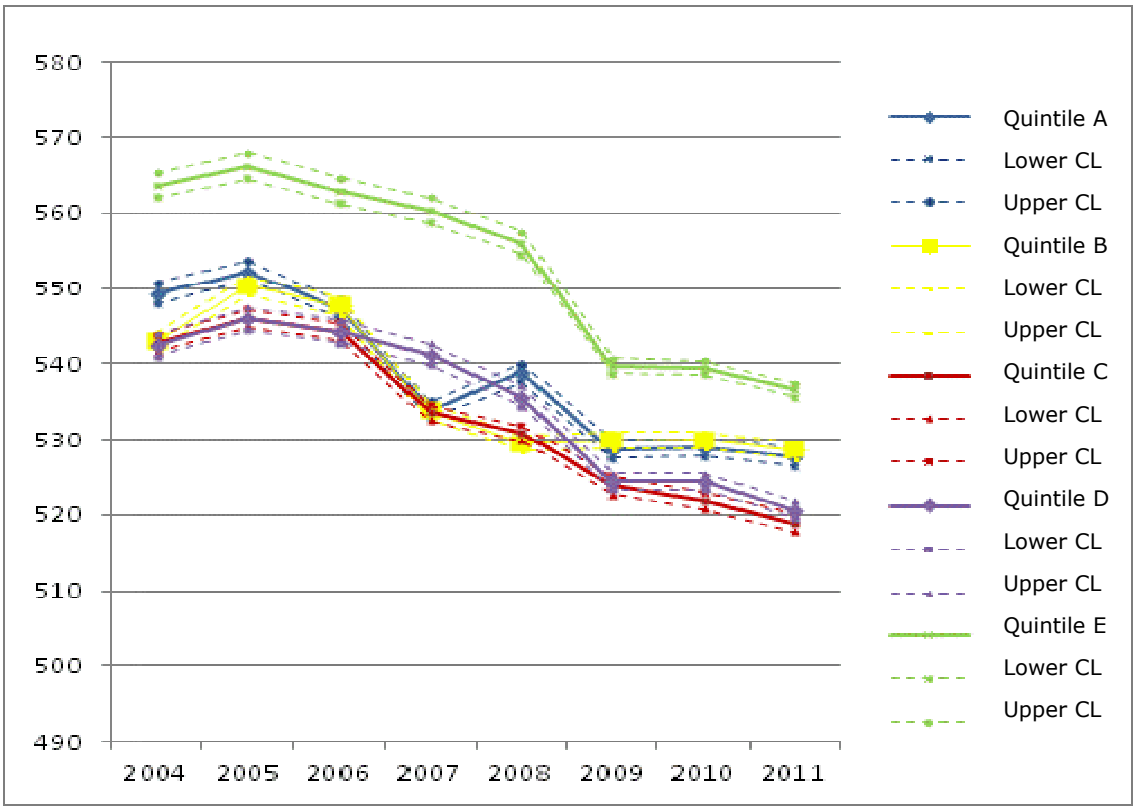


Figure 80: NEM by Socioeconomic Status

Appendix I. Summary Statistics for PSU Subtests by Subpopulations

Table 199: PSU Subtest by Gender

Year	PSU Subtest	Gender	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	Language and Communication	M	60773	490.51	124.44	0.50	489.52	491.49
2004	Language and Communication	F	67932	491.71	120.34	0.46	490.80	492.61
2005	Language and Communication	M	60905	492.90	114.43	0.46	491.99	493.81
2005	Language and Communication	F	67379	487.63	110.53	0.43	486.80	488.47
2006	Language and Communication	M	62672	490.54	114.84	0.46	489.64	491.44
2006	Language and Communication	F	70272	489.49	110.82	0.42	488.67	490.31
2007	Language and Communication	M	75347	493.02	114.13	0.42	492.20	493.83
2007	Language and Communication	F	87702	487.89	110.02	0.37	487.16	488.62
2008	Language and Communication	M	75838	497.18	113.16	0.41	496.37	497.98
2008	Language and Communication	F	89089	489.25	111.70	0.37	488.52	489.99
2009	Language and Communication	M	88289	500.48	115.29	0.39	499.72	501.25
2009	Language and Communication	F	97641	492.10	111.13	0.36	491.40	492.79
2010	Language and Communication	M	96380	499.11	113.58	0.37	498.39	499.82
2010	Language and Communication	F	104140	495.87	111.46	0.35	495.19	496.55
2011	Language and Communication	M	89935	493.75	114.05	0.38	493.01	494.50
2011	Language and Communication	F	97884	494.20	108.52	0.35	493.52	494.88
2004	Mathematics	M	60773	507.82	111.13	0.45	506.94	508.71
2004	Mathematics	F	67932	474.95	107.59	0.41	474.14	475.76
2005	Mathematics	M	60905	507.62	113.74	0.46	506.72	508.52
2005	Mathematics	F	67379	473.95	110.13	0.42	473.12	474.78
2006	Mathematics	M	62672	507.86	115.57	0.46	506.95	508.76
2006	Mathematics	F	70272	475.90	108.99	0.41	475.10	476.71
2007	Mathematics	M	75347	507.41	113.61	0.41	506.60	508.22
2007	Mathematics	F	87702	476.55	108.73	0.37	475.83	477.27
2008	Mathematics	M	75838	510.47	115.56	0.42	509.65	511.29
2008	Mathematics	F	89089	479.26	109.25	0.37	478.55	479.98
2009	Mathematics	M	88289	511.29	115.90	0.39	510.52	512.05
2009	Mathematics	F	97641	483.07	111.13	0.36	482.37	483.77
2010	Mathematics	M	96380	516.81	116.11	0.37	516.07	517.54
2010	Mathematics	F	104140	482.88	111.38	0.35	482.20	483.55
2011	Mathematics	M	89935	510.43	116.58	0.39	509.67	511.19
2011	Mathematics	F	97884	481.51	109.73	0.35	480.82	482.20

Year	PSU Subtest	Gender	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	Science	M	34195	497.42	103.37	0.56	496.33	498.52
2004	Science	F	33296	489.07	99.37	0.54	488.01	490.14
2005	Science	M	34565	501.75	113.68	0.61	500.56	502.95
2005	Science	F	33569	478.23	109.80	0.60	477.05	479.40
2006	Science	M	35910	499.18	115.41	0.61	497.98	500.37
2006	Science	F	36187	482.54	109.76	0.58	481.41	483.67
2007	Science	M	40437	504.99	114.34	0.57	503.87	506.10
2007	Science	F	42978	477.60	109.72	0.53	476.56	478.64
2008	Science	M	42358	510.83	111.99	0.54	509.77	511.90
2008	Science	F	46316	477.63	111.02	0.52	476.61	478.64
2009	Science	M	50076	511.60	116.69	0.52	510.58	512.62
2009	Science	F	53637	483.62	108.10	0.47	482.70	484.53
2010	Science	M	54120	514.65	111.83	0.48	513.71	515.59
2010	Science	F	57113	484.38	110.75	0.46	483.47	485.29
2011	Science	M	47568	509.82	113.72	0.52	508.80	510.84
2011	Science	F	52852	479.31	109.25	0.48	478.37	480.24
2004	History and Social Sciences	M	41412	509.42	110.84	0.54	508.35	510.49
2004	History and Social Sciences	F	50018	479.43	101.52	0.45	478.54	480.32
2005	History and Social Sciences	M	40033	505.66	112.31	0.56	504.56	506.76
2005	History and Social Sciences	F	48503	479.22	108.38	0.49	478.25	480.18
2006	History and Social Sciences	M	41664	504.44	112.70	0.55	503.36	505.53
2006	History and Social Sciences	F	49450	479.05	108.06	0.49	478.09	480.00
2007	History and Social Sciences	M	49517	505.50	111.80	0.50	504.51	506.48
2007	History and Social Sciences	F	60796	478.27	108.80	0.44	477.40	479.13
2008	History and Social Sciences	M	49197	508.38	112.52	0.51	507.38	509.37
2008	History and Social Sciences	F	60211	479.75	108.03	0.44	478.88	480.61
2009	History and Social Sciences	M	56829	507.37	114.62	0.48	506.43	508.32
2009	History and Social Sciences	F	64074	483.23	108.20	0.43	482.39	484.07
2010	History and Social Sciences	M	62300	510.09	112.46	0.45	509.21	510.98
2010	History and Social Sciences	F	67212	484.73	109.06	0.42	483.90	485.55
2011	History and Social Sciences	M	58240	504.09	112.61	0.47	503.18	505.01
2011	History and Social Sciences	F	62326	482.50	108.33	0.43	481.65	483.35
2004	NEM	M	59810	536.13	101.58	0.42	535.32	536.95
2004	NEM	F	67212	568.47	98.66	0.38	567.73	569.22

Year	PSU Subtest	Gender	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2005	NEM	M	58949	542.34	101.61	0.42	541.52	543.16
2005	NEM	F	66104	570.71	98.91	0.38	569.95	571.46
2006	NEM	M	60715	539.15	101.30	0.41	538.34	539.95
2006	NEM	F	68833	566.90	98.71	0.38	566.16	567.63
2007	NEM	M	72480	531.11	102.04	0.38	530.37	531.85
2007	NEM	F	85491	555.59	100.02	0.34	554.92	556.26
2008	NEM	M	73588	529.75	101.11	0.37	529.02	530.48
2008	NEM	F	87053	553.28	100.30	0.34	552.61	553.95
2009	NEM	M	85969	526.15	102.11	0.35	525.47	526.83
2009	NEM	F	95866	551.50	101.46	0.33	550.86	552.15
2010	NEM	M	93950	522.43	100.55	0.33	521.79	523.07
2010	NEM	F	102346	549.21	100.28	0.31	548.59	549.82
2011	NEM	M	88869	520.55	100.36	0.34	519.89	521.21
2011	NEM	F	97125	547.29	100.32	0.32	546.66	547.92
2004	LangComm and Mathematics	M	60773	499.16	109.49	0.44	498.29	500.03
2004	LangComm and Mathematics	F	67932	483.33	106.00	0.41	482.53	484.13
2005	LangComm and Mathematics	M	60905	500.26	106.88	0.43	499.41	501.11
2005	LangComm and Mathematics	F	67379	480.79	103.53	0.40	480.01	481.57
2006	LangComm and Mathematics	M	62672	499.20	107.89	0.43	498.35	500.04
2006	LangComm and Mathematics	F	70272	482.70	102.65	0.39	481.94	483.46
2007	LangComm and Mathematics	M	75347	500.21	106.72	0.39	499.45	500.97
2007	LangComm and Mathematics	F	87702	482.22	102.31	0.35	481.54	482.90
2008	LangComm and Mathematics	M	75838	503.82	107.80	0.39	503.06	504.59
2008	LangComm and Mathematics	F	89089	484.26	104.09	0.35	483.58	484.94
2009	LangComm and Mathematics	M	88289	505.89	109.26	0.37	505.16	506.61
2009	LangComm and Mathematics	F	97641	487.58	104.75	0.34	486.93	488.24
2010	LangComm and Mathematics	M	96380	507.96	108.60	0.35	507.27	508.64
2010	LangComm and Mathematics	F	104140	489.37	105.25	0.33	488.73	490.01
2011	LangComm and Mathematics	M	89935	502.09	108.16	0.36	501.38	502.80
2011	LangComm and Mathematics	F	97884	487.86	102.01	0.33	487.22	488.50

Table 200: PSU Subtest by School Type

Year	PSU Subtest	School Type	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	Language and Communication	Private	16543	584.95	118.24	0.92	583.14	586.75
2004	Language and Communication	Subsid.	54542	492.38	116.20	0.50	491.41	493.36
2004	Language and Communication	Munic.	56690	462.85	115.26	0.48	461.90	463.80
2005	Language and Communication	Private	15844	583.63	107.39	0.85	581.96	585.31
2005	Language and Communication	Subsid.	54967	490.33	105.55	0.45	489.45	491.21
2005	Language and Communication	Munic.	55302	465.29	106.39	0.45	464.40	466.17
2006	Language and Communication	Private	16406	587.69	105.96	0.83	586.06	589.31
2006	Language and Communication	Subsid.	58452	490.52	104.68	0.43	489.67	491.37
2006	Language and Communication	Munic.	56219	462.44	107.12	0.45	461.55	463.32
2007	Language and Communication	Private	17246	600.65	101.42	0.77	599.14	602.17
2007	Language and Communication	Subsid.	72769	493.14	104.44	0.39	492.38	493.90
2007	Language and Communication	Munic.	71083	461.51	105.00	0.39	460.74	462.28
2008	Language and Communication	Private	18555	600.46	100.70	0.74	599.01	601.91
2008	Language and Communication	Subsid.	77706	493.25	104.96	0.38	492.52	493.99
2008	Language and Communication	Munic.	67631	463.27	105.73	0.41	462.47	464.06
2009	Language and Communication	Private	20242	607.00	100.93	0.71	605.61	608.39
2009	Language and Communication	Subsid.	88217	496.58	105.13	0.35	495.89	497.28
2009	Language and Communication	Munic.	76410	466.18	106.88	0.39	465.42	466.93
2010	Language and Communication	Private	21477	608.26	97.89	0.67	606.95	609.57
2010	Language and Communication	Subsid.	96140	499.28	104.13	0.34	498.62	499.94
2010	Language and Communication	Munic.	81282	466.07	106.88	0.37	465.34	466.81
2011	Language and Communication	Private	20043	607.19	99.74	0.70	605.81	608.57
2011	Language and Communication	Subsid.	91859	494.44	102.75	0.34	493.77	495.10
2011	Language and Communication	Munic.	74156	463.06	104.42	0.38	462.30	463.81
2004	Mathematics	Private	16543	581.11	114.93	0.89	579.36	582.87
2004	Mathematics	Subsid.	54542	489.57	102.22	0.44	488.71	490.43
2004	Mathematics	Munic.	56690	465.44	102.83	0.43	464.60	466.29

Year	PSU Subtest	School Type	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2005	Mathematics	Private	15844	591.54	112.45	0.89	589.79	593.29
2005	Mathematics	Subsid.	54967	487.21	105.35	0.45	486.33	488.09
2005	Mathematics	Munic.	55302	466.15	104.74	0.45	465.28	467.02
2006	Mathematics	Private	16406	597.79	111.35	0.87	596.08	599.49
2006	Mathematics	Subsid.	58452	487.80	104.45	0.43	486.95	488.65
2006	Mathematics	Munic.	56219	465.25	104.56	0.44	464.38	466.11
2007	Mathematics	Private	17246	609.65	106.08	0.81	608.07	611.23
2007	Mathematics	Subsid.	72769	491.08	104.35	0.39	490.32	491.83
2007	Mathematics	Munic.	71083	463.19	102.10	0.38	462.44	463.94
2008	Mathematics	Private	18555	612.90	103.79	0.76	611.40	614.39
2008	Mathematics	Subsid.	77706	492.19	103.83	0.37	491.46	492.92
2008	Mathematics	Munic.	67631	463.12	104.54	0.40	462.33	463.91
2009	Mathematics	Private	20242	617.78	102.38	0.72	616.37	619.19
2009	Mathematics	Subsid.	88217	495.64	105.93	0.36	494.95	496.34
2009	Mathematics	Munic.	76410	465.62	105.15	0.38	464.88	466.37
2010	Mathematics	Private	21477	622.19	106.36	0.73	620.77	623.62
2010	Mathematics	Subsid.	96140	498.30	106.16	0.34	497.63	498.97
2010	Mathematics	Munic.	81282	468.19	105.39	0.37	467.46	468.91
2011	Mathematics	Private	20043	621.13	110.39	0.78	619.60	622.65
2011	Mathematics	Subsid.	91859	493.78	104.32	0.34	493.10	494.45
2011	Mathematics	Munic.	74156	464.04	103.12	0.38	463.30	464.78
2004	Science	Private	8901	568.87	100.54	1.07	566.78	570.95
2004	Science	Subsid.	29403	492.19	93.97	0.55	491.11	493.26
2004	Science	Munic.	28786	471.28	97.98	0.58	470.15	472.41
2005	Science	Private	8601	587.26	107.77	1.16	584.98	589.54
2005	Science	Subsid.	29741	487.79	103.61	0.60	486.62	488.97
2005	Science	Munic.	28863	465.50	106.91	0.63	464.27	466.74
2006	Science	Private	9161	585.01	107.80	1.13	582.80	587.22
2006	Science	Subsid.	32289	488.89	104.02	0.58	487.76	490.03
2006	Science	Munic.	29863	465.73	108.59	0.63	464.50	466.97
2007	Science	Private	9700	596.31	102.25	1.04	594.27	598.34
2007	Science	Subsid.	38646	490.80	103.43	0.53	489.77	491.83
2007	Science	Munic.	34171	462.53	108.21	0.59	461.38	463.67
2008	Science	Private	10533	598.06	101.29	0.99	596.13	600.00
2008	Science	Subsid.	42877	491.07	103.61	0.50	490.09	492.05
2008	Science	Munic.	34717	465.39	108.48	0.58	464.25	466.53
2009	Science	Private	11864	602.67	101.97	0.94	600.83	604.50
2009	Science	Subsid.	50362	496.32	104.69	0.47	495.40	497.23
2009	Science	Munic.	40918	467.80	108.34	0.54	466.75	468.85
2010	Science	Private	12822	602.29	101.75	0.90	600.53	604.05
2010	Science	Subsid.	55136	497.93	103.23	0.44	497.07	498.79
2010	Science	Munic.	42429	470.08	108.67	0.53	469.05	471.12
2011	Science	Private	11567	602.25	104.11	0.97	600.35	604.15
2011	Science	Subsid.	50906	492.01	103.08	0.46	491.12	492.91
2011	Science	Munic.	37045	463.13	106.72	0.55	462.04	464.22
2004	History and Social Sciences	Private	10719	575.99	113.61	1.10	573.84	578.14

Year	PSU Subtest	School Type	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	History and Social Sciences	Subsid.	38368	491.40	100.82	0.51	490.39	492.41
2004	History and Social Sciences	Munic.	41664	473.24	100.34	0.49	472.27	474.20
2005	History and Social Sciences	Private	10165	577.46	110.79	1.10	575.30	579.61
2005	History and Social Sciences	Subsid.	37464	491.20	105.38	0.54	490.13	492.27
2005	History and Social Sciences	Munic.	39364	470.31	105.56	0.53	469.27	471.35
2006	History and Social Sciences	Private	10483	579.56	109.72	1.07	577.46	581.66
2006	History and Social Sciences	Subsid.	39700	490.64	104.30	0.52	489.61	491.66
2006	History and Social Sciences	Munic.	39602	468.14	106.16	0.53	467.10	469.19
2007	History and Social Sciences	Private	10963	592.43	105.29	1.01	590.46	594.40
2007	History and Social Sciences	Subsid.	48392	493.15	104.77	0.48	492.22	494.08
2007	History and Social Sciences	Munic.	49603	466.06	105.05	0.47	465.13	466.98
2008	History and Social Sciences	Private	11505	593.92	103.21	0.96	592.03	595.81
2008	History and Social Sciences	Subsid.	50618	493.81	103.91	0.46	492.90	494.71
2008	History and Social Sciences	Munic.	46571	466.37	105.77	0.49	465.41	467.33
2009	History and Social Sciences	Private	12195	598.38	104.29	0.94	596.53	600.23
2009	History and Social Sciences	Subsid.	56279	496.59	104.94	0.44	495.72	497.46
2009	History and Social Sciences	Munic.	51719	467.71	106.42	0.47	466.79	468.62
2010	History and Social Sciences	Private	12708	600.62	103.20	0.92	598.83	602.41
2010	History and Social Sciences	Subsid.	60647	500.03	104.19	0.42	499.20	500.86
2010	History and Social Sciences	Munic.	55103	469.29	106.43	0.45	468.40	470.18
2011	History and Social Sciences	Private	11760	600.97	105.96	0.98	599.06	602.89
2011	History and Social Sciences	Subsid.	57730	494.32	103.40	0.43	493.47	495.16
2011	History and Social Sciences	Munic.	49989	465.77	104.83	0.47	464.85	466.69
2004	NEM	Private	16335	588.99	107.69	0.84	587.34	590.65
2004	NEM	Subsid.	53897	551.20	100.20	0.43	550.36	552.05
2004	NEM	Munic.	56052	545.75	97.92	0.41	544.94	546.56
2005	NEM	Private	15515	596.13	106.66	0.86	594.45	597.81
2005	NEM	Subsid.	54198	552.02	100.20	0.43	551.18	552.87
2005	NEM	Munic.	54720	552.27	97.84	0.42	551.45	553.09

Year	PSU Subtest	School Type	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2006	NEM	Private	15771	594.88	106.87	0.85	593.22	596.55
2006	NEM	Subsid.	57550	548.38	100.12	0.42	547.56	549.19
2006	NEM	Munic.	55560	548.73	97.00	0.41	547.92	549.54
2007	NEM	Private	16543	595.70	106.61	0.83	594.08	597.33
2007	NEM	Subsid.	70987	540.54	101.05	0.38	539.80	541.29
2007	NEM	Munic.	69661	536.76	97.45	0.37	536.03	537.48
2008	NEM	Private	17728	590.48	106.77	0.80	588.90	592.05
2008	NEM	Subsid.	75661	537.50	100.20	0.36	536.79	538.21
2008	NEM	Munic.	66405	536.26	97.52	0.38	535.51	537.00
2009	NEM	Private	19372	590.90	107.41	0.77	589.39	592.42
2009	NEM	Subsid.	86280	534.39	101.20	0.34	533.72	535.07
2009	NEM	Munic.	75315	532.84	98.95	0.36	532.13	533.54
2010	NEM	Private	20567	589.63	105.75	0.74	588.18	591.08
2010	NEM	Subsid.	94066	531.75	100.14	0.33	531.11	532.39
2010	NEM	Munic.	80282	529.31	97.19	0.34	528.64	529.99
2011	NEM	Private	19785	592.06	106.97	0.76	590.57	593.55
2011	NEM	Subsid.	91125	528.23	99.90	0.33	527.58	528.88
2011	NEM	Munic.	73628	527.92	96.23	0.35	527.23	528.62
2004	LangComm and Mathematics	Private	16543	583.03	107.66	0.84	581.39	584.67
2004	LangComm and Mathematics	Subsid.	54542	490.98	100.36	0.43	490.14	491.82
2004	LangComm and Mathematics	Munic.	56690	464.15	100.03	0.42	463.32	464.97
2005	LangComm and Mathematics	Private	15844	587.59	102.22	0.81	586.00	589.18
2005	LangComm and Mathematics	Subsid.	54967	488.77	97.87	0.42	487.95	489.59
2005	LangComm and Mathematics	Munic.	55302	465.72	97.89	0.42	464.90	466.53
2006	LangComm and Mathematics	Private	16406	592.74	100.95	0.79	591.19	594.28
2006	LangComm and Mathematics	Subsid.	58452	489.16	96.58	0.40	488.38	489.94
2006	LangComm and Mathematics	Munic.	56219	463.84	97.65	0.41	463.03	464.65
2007	LangComm and Mathematics	Private	17246	605.15	96.30	0.73	603.71	606.59
2007	LangComm and Mathematics	Subsid.	72769	492.11	96.77	0.36	491.40	492.81
2007	LangComm and Mathematics	Munic.	71083	462.35	95.42	0.36	461.65	463.05
2008	LangComm and Mathematics	Private	18555	606.68	95.01	0.70	605.31	608.05
2008	LangComm and Mathematics	Subsid.	77706	492.73	97.42	0.35	492.04	493.41
2008	LangComm and Mathematics	Munic.	67631	463.19	97.86	0.38	462.45	463.93
2009	LangComm and Mathematics	Private	20242	612.39	94.67	0.67	611.09	613.70

Year	PSU Subtest	School Type	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2009	LangComm and Mathematics	Subsid.	88217	496.11	98.71	0.33	495.46	496.76
2009	LangComm and Mathematics	Munic.	76410	465.90	98.89	0.36	465.20	466.60
2010	LangComm and Mathematics	Private	21477	615.23	94.72	0.65	613.96	616.49
2010	LangComm and Mathematics	Subsid.	96140	498.79	98.30	0.32	498.17	499.41
2010	LangComm and Mathematics	Munic.	81282	467.13	99.16	0.35	466.45	467.81
2011	LangComm and Mathematics	Private	20043	614.16	97.37	0.69	612.81	615.51
2011	LangComm and Mathematics	Subsid.	91859	494.11	95.65	0.32	493.49	494.73
2011	LangComm and Mathematics	Munic.	74156	463.55	95.60	0.35	462.86	464.24

Table 201: PSU Subtest by School Type

Year	PSU Subtest	Ed. Branch	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	Language and Communication	Scientific	94293	508.12	124.19	0.40	507.33	508.91
2004	Language and Communication	Technical	33475	443.80	103.03	0.56	442.69	444.90
2005	Language and Communication	Scientific	92028	507.94	113.60	0.37	507.21	508.68
2005	Language and Communication	Technical	33965	445.53	95.26	0.52	444.52	446.55
2006	Language and Communication	Scientific	96736	508.85	113.39	0.36	508.14	509.57
2006	Language and Communication	Technical	34236	439.32	93.96	0.51	438.33	440.32
2007	Language and Communication	Scientific	109098	514.85	112.80	0.34	514.18	515.52
2007	Language and Communication	Technical	51885	440.02	91.90	0.40	439.23	440.81
2008	Language and Communication	Scientific	114557	515.64	113.11	0.33	514.99	516.30
2008	Language and Communication	Technical	49323	440.49	91.59	0.41	439.68	441.30
2009	Language and Communication	Scientific	127929	519.58	114.29	0.32	518.95	520.20
2009	Language and Communication	Technical	56927	443.37	91.01	0.38	442.62	444.12
2010	Language and Communication	Scientific	137998	521.85	112.88	0.30	521.25	522.44
2010	Language and Communication	Technical	60896	442.26	90.14	0.37	441.54	442.97
2011	Language and Communication	Scientific	127898	517.74	113.16	0.32	517.12	518.36
2011	Language and Communication	Technical	58152	442.02	86.84	0.36	441.32	442.73
2004	Mathematics	Scientific	94293	505.88	113.33	0.37	505.15	506.60
2004	Mathematics	Technical	33475	448.04	89.08	0.49	447.08	448.99
2005	Mathematics	Scientific	92028	509.05	115.43	0.38	508.30	509.79
2005	Mathematics	Technical	33965	442.65	90.03	0.49	441.69	443.60
2006	Mathematics	Scientific	96736	510.00	115.18	0.37	509.27	510.72
2006	Mathematics	Technical	34236	440.96	89.74	0.48	440.01	441.91
2007	Mathematics	Scientific	109098	515.56	114.82	0.35	514.88	516.24
2007	Mathematics	Technical	51885	440.95	86.68	0.38	440.20	441.69
2008	Mathematics	Scientific	114557	517.99	114.63	0.34	517.32	518.65
2008	Mathematics	Technical	49323	437.86	87.40	0.39	437.09	438.63
2009	Mathematics	Scientific	127929	522.56	115.25	0.32	521.93	523.19
2009	Mathematics	Technical	56927	438.31	87.78	0.37	437.59	439.04
2010	Mathematics	Scientific	137998	524.27	116.84	0.31	523.65	524.88
2010	Mathematics	Technical	60896	442.95	87.77	0.36	442.26	443.65
2011	Mathematics	Scientific	127898	520.24	117.55	0.33	519.60	520.88
2011	Mathematics	Technical	58152	441.56	83.73	0.35	440.88	442.24
2004	Science	Scientific	53934	505.88	101.44	0.44	505.02	506.73

Year	PSU Subtest	Ed. Branch	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	Science	Technical	13155	442.21	84.05	0.73	440.77	443.64
2005	Science	Scientific	53774	504.74	113.13	0.49	503.78	505.70
2005	Science	Technical	13374	435.77	89.85	0.78	434.25	437.29
2006	Science	Scientific	57844	504.65	113.34	0.47	503.73	505.58
2006	Science	Technical	13418	435.26	91.39	0.79	433.72	436.81
2007	Science	Scientific	63916	509.67	111.98	0.44	508.81	510.54
2007	Science	Technical	18555	429.01	91.30	0.67	427.70	430.32
2008	Science	Scientific	69619	511.11	111.96	0.42	510.28	511.94
2008	Science	Technical	18500	428.44	89.13	0.66	427.16	429.73
2009	Science	Scientific	80641	514.89	113.09	0.40	514.11	515.67
2009	Science	Technical	22496	433.97	88.52	0.59	432.82	435.13
2010	Science	Scientific	86971	516.91	111.74	0.38	516.16	517.65
2010	Science	Technical	23414	434.13	88.02	0.58	433.00	435.26
2011	Science	Scientific	78415	512.33	112.25	0.40	511.55	513.12
2011	Science	Technical	21099	426.22	83.75	0.58	425.09	427.35
2004	History and Social Sciences	Scientific	64022	505.67	110.16	0.44	504.81	506.52
2004	History and Social Sciences	Technical	26723	462.83	92.07	0.56	461.73	463.94
2005	History and Social Sciences	Scientific	60625	505.93	112.98	0.46	505.03	506.83
2005	History and Social Sciences	Technical	26293	459.34	99.04	0.61	458.14	460.53
2006	History and Social Sciences	Scientific	62882	506.15	113.14	0.45	505.27	507.04
2006	History and Social Sciences	Technical	26827	455.87	97.37	0.59	454.70	457.03
2007	History and Social Sciences	Scientific	68831	512.71	112.47	0.43	511.87	513.55
2007	History and Social Sciences	Technical	40041	453.25	97.78	0.49	452.29	454.21
2008	History and Social Sciences	Scientific	70752	513.89	112.93	0.42	513.06	514.73
2008	History and Social Sciences	Technical	37935	453.04	95.56	0.49	452.08	454.00
2009	History and Social Sciences	Scientific	77273	518.77	113.33	0.41	517.98	519.57
2009	History and Social Sciences	Technical	42912	450.77	94.95	0.46	449.87	451.67
2010	History and Social Sciences	Scientific	82310	520.67	113.05	0.39	519.89	521.44
2010	History and Social Sciences	Technical	46144	454.22	94.79	0.44	453.36	455.09
2011	History and Social Sciences	Scientific	76028	518.11	112.88	0.41	517.31	518.91
2011	History and Social Sciences	Technical	43448	448.70	92.40	0.44	447.83	449.57
2004	NEM	Scientific	93225	559.61	102.71	0.34	558.95	560.27
2004	NEM	Technical	33053	536.94	94.65	0.52	535.92	537.96
2005	NEM	Scientific	90668	562.21	103.04	0.34	561.54	562.88

Year	PSU Subtest	Ed. Branch	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2005	NEM	Technical	33646	545.54	94.47	0.52	544.53	546.55
2006	NEM	Scientific	94912	558.54	102.82	0.33	557.89	559.20
2006	NEM	Technical	33864	542.34	93.86	0.51	541.34	543.34
2007	NEM	Scientific	106275	551.99	104.48	0.32	551.36	552.62
2007	NEM	Technical	50801	529.51	93.53	0.41	528.69	530.32
2008	NEM	Scientific	111490	549.39	103.92	0.31	548.78	550.00
2008	NEM	Technical	48295	527.80	93.09	0.42	526.97	528.63
2009	NEM	Scientific	124680	548.23	105.42	0.30	547.64	548.81
2009	NEM	Technical	56274	521.14	93.03	0.39	520.37	521.90
2010	NEM	Scientific	134303	546.10	103.57	0.28	545.55	546.65
2010	NEM	Technical	60607	516.37	92.48	0.38	515.63	517.11
2011	NEM	Scientific	126530	544.71	104.21	0.29	544.14	545.29
2011	NEM	Technical	58000	513.67	90.71	0.38	512.93	514.41
2004	LangComm and Mathematics	Scientific	94293	507.00	110.36	0.36	506.29	507.70
2004	LangComm and Mathematics	Technical	33475	445.92	86.21	0.47	444.99	446.84
2005	LangComm and Mathematics	Scientific	92028	508.50	107.44	0.35	507.80	509.19
2005	LangComm and Mathematics	Technical	33965	444.09	83.91	0.46	443.20	444.98
2006	LangComm and Mathematics	Scientific	96736	509.42	106.91	0.34	508.75	510.10
2006	LangComm and Mathematics	Technical	34236	440.14	82.33	0.44	439.27	441.01
2007	LangComm and Mathematics	Scientific	109098	515.20	106.86	0.32	514.57	515.84
2007	LangComm and Mathematics	Technical	51885	440.48	79.60	0.35	439.80	441.17
2008	LangComm and Mathematics	Scientific	114557	516.82	107.43	0.32	516.19	517.44
2008	LangComm and Mathematics	Technical	49323	439.17	80.86	0.36	438.46	439.89
2009	LangComm and Mathematics	Scientific	127929	521.07	108.56	0.30	520.47	521.66
2009	LangComm and Mathematics	Technical	56927	440.84	80.66	0.34	440.18	441.50
2010	LangComm and Mathematics	Scientific	137998	523.06	108.58	0.29	522.48	523.63
2010	LangComm and Mathematics	Technical	60896	442.61	80.35	0.33	441.97	443.24
2011	LangComm and Mathematics	Scientific	127898	518.99	108.29	0.30	518.40	519.59
2011	LangComm and Mathematics	Technical	58152	441.79	75.09	0.31	441.18	442.40

Table 202: PSU Subtest by Region

Year	PSU Subtest	Region	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	Language and Communication	C	71000	497.49	125.25	0.47	496.57	498.41
2004	Language and Communication	N	15533	472.00	114.90	0.92	470.19	473.81
2004	Language and Communication	S	42055	487.33	118.95	0.58	486.19	488.46
2005	Language and Communication	C	67793	497.11	116.14	0.45	496.24	497.99
2005	Language and Communication	N	16290	474.11	105.16	0.82	472.50	475.72
2005	Language and Communication	S	42713	487.51	108.19	0.52	486.48	488.54
2006	Language and Communication	C	70668	496.68	116.61	0.44	495.82	497.54
2006	Language and Communication	N	16586	470.88	104.83	0.81	469.29	472.48
2006	Language and Communication	S	44534	488.04	108.48	0.51	487.03	489.05
2007	Language and Communication	C	86450	495.90	115.79	0.39	495.13	496.68
2007	Language and Communication	N	18744	476.93	104.28	0.76	475.44	478.43
2007	Language and Communication	S	56760	487.06	108.07	0.45	486.17	487.95
2008	Language and Communication	C	86774	499.48	116.24	0.39	498.70	500.25
2008	Language and Communication	N	19191	478.99	104.04	0.75	477.52	480.46
2008	Language and Communication	S	58889	487.65	108.63	0.45	486.77	488.53
2009	Language and Communication	C	95931	501.98	116.60	0.38	501.24	502.72
2009	Language and Communication	N	21391	480.33	105.52	0.72	478.92	481.74
2009	Language and Communication	S	68501	492.66	110.01	0.42	491.84	493.48
2010	Language and Communication	C	106097	504.38	116.70	0.36	503.67	505.08
2010	Language and Communication	N	22878	483.78	104.67	0.69	482.43	485.14
2010	Language and Communication	S	71424	491.38	107.70	0.40	490.59	492.17
2011	Language and Communication	C	100383	503.52	114.68	0.36	502.81	504.23
2011	Language and Communication	N	20955	478.95	106.15	0.73	477.51	480.39
2011	Language and Communication	S	66331	484.23	105.94	0.41	483.42	485.03
2004	Mathematics	C	71000	491.43	114.61	0.43	490.59	492.27
2004	Mathematics	N	15533	482.49	101.59	0.82	480.89	484.09
2004	Mathematics	S	42055	491.59	106.31	0.52	490.57	492.61

Year	PSU Subtest	Region	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2005	Mathematics	C	67793	493.44	117.19	0.45	492.56	494.32
2005	Mathematics	N	16290	480.63	103.89	0.81	479.03	482.23
2005	Mathematics	S	42713	490.55	109.48	0.53	489.52	491.59
2006	Mathematics	C	70668	493.73	118.11	0.44	492.86	494.60
2006	Mathematics	N	16586	480.67	104.23	0.81	479.09	482.26
2006	Mathematics	S	44534	492.29	108.20	0.51	491.29	493.30
2007	Mathematics	C	86450	493.43	116.26	0.40	492.65	494.20
2007	Mathematics	N	18744	484.45	102.53	0.75	482.98	485.91
2007	Mathematics	S	56760	490.20	108.40	0.45	489.31	491.09
2008	Mathematics	C	86774	498.05	117.33	0.40	497.27	498.83
2008	Mathematics	N	19191	487.08	103.13	0.74	485.62	488.54
2008	Mathematics	S	58889	489.13	110.00	0.45	488.24	490.01
2009	Mathematics	C	95931	498.96	118.26	0.38	498.21	499.71
2009	Mathematics	N	21391	486.86	105.32	0.72	485.45	488.27
2009	Mathematics	S	68501	495.87	111.11	0.42	495.04	496.70
2010	Mathematics	C	106097	503.70	119.55	0.37	502.98	504.42
2010	Mathematics	N	22878	491.53	105.92	0.70	490.16	492.91
2010	Mathematics	S	71424	494.84	110.31	0.41	494.04	495.65
2011	Mathematics	C	100383	502.42	118.30	0.37	501.68	503.15
2011	Mathematics	N	20955	487.03	107.64	0.74	485.57	488.48
2011	Mathematics	S	66331	487.22	108.41	0.42	486.40	488.05
2004	Science	C	33962	499.47	105.64	0.57	498.35	500.60
2004	Science	N	8448	477.19	94.19	1.02	475.18	479.20
2004	Science	S	25007	490.14	97.22	0.61	488.93	491.34
2005	Science	C	33089	499.55	117.13	0.64	498.29	500.82
2005	Science	N	8955	474.36	103.04	1.09	472.22	476.49
2005	Science	S	25479	485.08	108.01	0.68	483.76	486.41
2006	Science	C	34898	501.43	117.15	0.63	500.20	502.66
2006	Science	N	9548	471.35	104.24	1.07	469.26	473.44
2006	Science	S	27182	485.31	108.71	0.66	484.01	486.60
2007	Science	C	39999	501.96	116.48	0.58	500.82	503.10
2007	Science	N	10299	474.03	101.60	1.00	472.07	475.99
2007	Science	S	32635	483.69	110.22	0.61	482.49	484.88
2008	Science	C	42450	504.00	116.57	0.57	502.89	505.11
2008	Science	N	10865	480.61	101.64	0.98	478.70	482.52
2008	Science	S	35310	484.74	110.03	0.59	483.59	485.89
2009	Science	C	48123	507.75	116.01	0.53	506.71	508.78
2009	Science	N	12541	480.22	105.42	0.94	478.37	482.06
2009	Science	S	42975	490.10	111.01	0.54	489.05	491.15
2010	Science	C	53441	510.93	116.52	0.50	509.95	511.92
2010	Science	N	13462	485.13	101.72	0.88	483.41	486.85
2010	Science	S	44254	488.99	108.65	0.52	487.97	490.00
2011	Science	C	49033	507.19	116.25	0.53	506.16	508.22
2011	Science	N	11994	481.46	105.21	0.96	479.58	483.35
2011	Science	S	39303	480.66	107.64	0.54	479.60	481.73
2004	History and Social Sciences	C	49945	500.27	111.95	0.50	499.29	501.26

Year	PSU Subtest	Region	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	History and Social Sciences	N	10937	473.50	96.28	0.92	471.70	475.31
2004	History and Social Sciences	S	30494	488.08	100.69	0.58	486.95	489.21
2005	History and Social Sciences	C	46299	499.24	115.68	0.54	498.19	500.29
2005	History and Social Sciences	N	11152	473.08	101.91	0.97	471.19	474.98
2005	History and Social Sciences	S	30039	487.34	105.67	0.61	486.15	488.54
2006	History and Social Sciences	C	48422	498.58	116.14	0.53	497.54	499.61
2006	History and Social Sciences	N	11021	468.86	101.98	0.97	466.96	470.77
2006	History and Social Sciences	S	30857	487.32	104.44	0.59	486.16	488.49
2007	History and Social Sciences	C	58910	495.49	116.14	0.48	494.55	496.43
2007	History and Social Sciences	N	12371	476.39	102.49	0.92	474.58	478.19
2007	History and Social Sciences	S	38276	488.21	104.98	0.54	487.16	489.26
2008	History and Social Sciences	C	58085	500.39	115.15	0.48	499.45	501.33
2008	History and Social Sciences	N	12359	477.01	103.37	0.93	475.19	478.83
2008	History and Social Sciences	S	38930	485.93	105.93	0.54	484.88	486.98
2009	History and Social Sciences	C	63308	500.75	116.41	0.46	499.84	501.65
2009	History and Social Sciences	N	13561	476.96	104.03	0.89	475.21	478.71
2009	History and Social Sciences	S	43984	491.08	106.82	0.51	490.08	492.08
2010	History and Social Sciences	C	69238	504.51	115.73	0.44	503.65	505.37
2010	History and Social Sciences	N	14189	480.02	105.11	0.88	478.29	481.75
2010	History and Social Sciences	S	46026	490.63	105.62	0.49	489.66	491.59
2011	History and Social Sciences	C	65393	502.41	115.08	0.45	501.53	503.30
2011	History and Social Sciences	N	12836	475.66	104.12	0.92	473.86	477.46
2011	History and Social Sciences	S	42265	483.43	104.82	0.51	482.43	484.43
2004	NEM	C	69878	546.38	101.25	0.38	545.63	547.13
2004	NEM	N	15393	559.18	103.11	0.83	557.56	560.81
2004	NEM	S	41713	562.49	99.93	0.49	561.53	563.45
2005	NEM	C	66601	551.29	101.10	0.39	550.52	552.06
2005	NEM	N	16132	562.23	102.21	0.80	560.65	563.80
2005	NEM	S	42291	564.93	100.29	0.49	563.97	565.88

Year	PSU Subtest	Region	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2006	NEM	C	69093	548.20	101.31	0.39	547.45	548.96
2006	NEM	N	16413	557.16	102.12	0.80	555.60	558.72
2006	NEM	S	44020	561.58	99.17	0.47	560.65	562.51
2007	NEM	C	83845	537.23	102.37	0.35	536.54	537.93
2007	NEM	N	18331	550.12	102.08	0.75	548.64	551.60
2007	NEM	S	55772	553.14	99.70	0.42	552.31	553.97
2008	NEM	C	83700	535.13	102.20	0.35	534.44	535.82
2008	NEM	N	18791	547.26	102.10	0.74	545.80	548.72
2008	NEM	S	58123	551.54	99.02	0.41	550.74	552.35
2009	NEM	C	93221	530.43	102.71	0.34	529.77	531.09
2009	NEM	N	21051	543.08	102.09	0.70	541.70	544.46
2009	NEM	S	67527	550.92	101.25	0.39	550.16	551.68
2010	NEM	C	103068	528.51	102.13	0.32	527.89	529.14
2010	NEM	N	22573	542.80	101.42	0.68	541.48	544.12
2010	NEM	S	70612	545.80	99.06	0.37	545.07	546.54
2011	NEM	C	99221	528.34	102.40	0.33	527.70	528.98
2011	NEM	N	20830	539.95	101.20	0.70	538.58	541.33
2011	NEM	S	65906	542.08	98.80	0.38	541.33	542.84
2004	LangComm and Mathematics	C	71000	494.46	111.66	0.42	493.64	495.28
2004	LangComm and Mathematics	N	15533	477.25	99.47	0.80	475.68	478.81
2004	LangComm and Mathematics	S	42055	489.46	104.03	0.51	488.46	490.45
2005	LangComm and Mathematics	C	67793	495.28	109.61	0.42	494.45	496.10
2005	LangComm and Mathematics	N	16290	477.37	97.00	0.76	475.88	478.86
2005	LangComm and Mathematics	S	42713	489.03	101.55	0.49	488.07	489.99
2006	LangComm and Mathematics	C	70668	495.20	109.99	0.41	494.39	496.01
2006	LangComm and Mathematics	N	16586	475.78	96.58	0.75	474.31	477.25
2006	LangComm and Mathematics	S	44534	490.16	100.74	0.48	489.23	491.10
2007	LangComm and Mathematics	C	86450	494.67	108.91	0.37	493.94	495.39
2007	LangComm and Mathematics	N	18744	480.69	95.83	0.70	479.32	482.06
2007	LangComm and Mathematics	S	56760	488.63	100.87	0.42	487.80	489.46
2008	LangComm and Mathematics	C	86774	498.77	110.30	0.37	498.03	499.50
2008	LangComm and Mathematics	N	19191	483.03	96.75	0.70	481.67	484.40
2008	LangComm and Mathematics	S	58889	488.39	102.61	0.42	487.56	489.22
2009	LangComm and Mathematics	C	95931	500.47	111.04	0.36	499.77	501.17

Year	PSU Subtest	Region	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2009	LangComm and Mathematics	N	21391	483.60	98.72	0.67	482.27	484.92
2009	LangComm and Mathematics	S	68501	494.27	104.13	0.40	493.49	495.05
2010	LangComm and Mathematics	C	106097	504.04	111.81	0.34	503.36	504.71
2010	LangComm and Mathematics	N	22878	487.66	98.40	0.65	486.38	488.93
2010	LangComm and Mathematics	S	71424	493.11	102.48	0.38	492.36	493.86
2011	LangComm and Mathematics	C	100383	502.97	109.24	0.34	502.29	503.64
2011	LangComm and Mathematics	N	20955	482.99	99.25	0.69	481.64	484.33
2011	LangComm and Mathematics	S	66331	485.72	99.70	0.39	484.97	486.48

Table 203: PSU Subtest by SES Quintile

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2004	Language and Communication	A	17895	434.93	107.90	0.81	433.35	436.52
2004	Language and Communication	B	25316	454.20	109.54	0.69	452.85	455.55
2004	Language and Communication	C	27072	477.77	111.82	0.68	476.43	479.10
2004	Language and Communication	D	17482	494.78	114.78	0.87	493.08	496.49
2004	Language and Communication	E	16107	537.26	116.81	0.92	535.45	539.06
2005	Language and Communication	A	22425	447.82	107.14	0.72	446.42	449.22
2005	Language and Communication	B	25901	457.41	101.54	0.63	456.17	458.64
2005	Language and Communication	C	29134	477.19	100.76	0.59	476.03	478.35
2005	Language and Communication	D	16947	494.21	105.83	0.81	492.61	495.80
2005	Language and Communication	E	13586	536.24	104.39	0.90	534.49	538.00
2006	Language and Communication	A	27186	451.72	109.64	0.66	450.42	453.03
2006	Language and Communication	B	25232	457.37	102.85	0.65	456.10	458.64
2006	Language and Communication	C	29328	476.45	101.58	0.59	475.28	477.61
2006	Language and Communication	D	17191	495.67	102.65	0.78	494.14	497.21
2006	Language and Communication	E	14030	537.04	104.05	0.88	535.32	538.76
2007	Language and Communication	A	33945	445.24	105.57	0.57	444.11	446.36
2007	Language and Communication	B	35105	455.57	99.88	0.53	454.52	456.61
2007	Language and Communication	C	37609	478.89	99.30	0.51	477.88	479.89
2007	Language and Communication	D	19325	507.56	100.93	0.73	506.13	508.98
2007	Language and Communication	E	14872	549.54	100.56	0.82	547.93	551.16
2008	Language and Communication	A	28410	448.93	110.59	0.66	447.65	450.22
2008	Language and Communication	B	34702	452.11	99.04	0.53	451.07	453.15
2008	Language and Communication	C	38608	479.57	100.25	0.51	478.57	480.57
2008	Language and Communication	D	22965	503.84	100.18	0.66	502.54	505.13
2008	Language and Communication	E	17180	545.36	100.86	0.77	543.86	546.87
2009	Language and	A	25247	438.25	107.51	0.68	436.92	439.58

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
	Communication							
2009	Language and Communication	B	30079	454.66	103.56	0.60	453.49	455.83
2009	Language and Communication	C	26798	465.53	99.05	0.61	464.34	466.71
2009	Language and Communication	D	32151	481.97	99.31	0.55	480.88	483.05
2009	Language and Communication	E	42848	526.76	101.52	0.49	525.80	527.73
2010	Language and Communication	A	27710	444.23	107.98	0.65	442.96	445.50
2010	Language and Communication	B	32913	459.29	106.34	0.59	458.14	460.44
2010	Language and Communication	C	28840	468.02	99.92	0.59	466.86	469.17
2010	Language and Communication	D	36217	485.74	100.22	0.53	484.71	486.78
2010	Language and Communication	E	48798	530.89	100.90	0.46	529.99	531.78
2011	Language and Communication	A	25163	442.35	106.60	0.67	441.04	443.67
2011	Language and Communication	B	29934	458.59	104.94	0.61	457.40	459.78
2011	Language and Communication	C	26078	463.74	97.17	0.60	462.56	464.92
2011	Language and Communication	D	34904	480.08	98.39	0.53	479.05	481.12
2011	Language and Communication	E	47746	523.55	100.40	0.46	522.65	524.45
2004	Mathematics	A	17895	439.77	97.10	0.73	438.34	441.19
2004	Mathematics	B	25316	456.50	97.34	0.61	455.30	457.70
2004	Mathematics	C	27072	476.93	97.41	0.59	475.77	478.09
2004	Mathematics	D	17482	491.84	100.98	0.76	490.35	493.34
2004	Mathematics	E	16107	532.29	106.23	0.84	530.65	533.93
2005	Mathematics	A	22425	448.48	104.49	0.70	447.11	449.84
2005	Mathematics	B	25901	456.09	100.22	0.62	454.87	457.31
2005	Mathematics	C	29134	473.87	101.21	0.59	472.71	475.03
2005	Mathematics	D	16947	493.70	104.40	0.80	492.13	495.27
2005	Mathematics	E	13586	537.88	106.24	0.91	536.09	539.66
2006	Mathematics	A	27186	456.11	107.29	0.65	454.84	457.39
2006	Mathematics	B	25232	458.21	101.93	0.64	456.95	459.47
2006	Mathematics	C	29328	474.07	100.61	0.59	472.92	475.22
2006	Mathematics	D	17191	492.52	103.90	0.79	490.97	494.08
2006	Mathematics	E	14030	537.78	106.92	0.90	536.01	539.55
2007	Mathematics	A	33945	449.10	102.27	0.56	448.01	450.19
2007	Mathematics	B	35105	456.49	98.02	0.52	455.46	457.51
2007	Mathematics	C	37609	476.23	99.57	0.51	475.22	477.23
2007	Mathematics	D	19325	505.85	102.64	0.74	504.40	507.30
2007	Mathematics	E	14872	549.19	102.99	0.84	547.54	550.85
2008	Mathematics	A	28410	449.29	110.59	0.66	448.01	450.58

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2008	Mathematics	B	34702	450.48	97.12	0.52	449.46	451.50
2008	Mathematics	C	38608	477.85	99.28	0.51	476.86	478.85
2008	Mathematics	D	22965	503.61	99.85	0.66	502.32	504.90
2008	Mathematics	E	17180	549.29	101.10	0.77	547.77	550.80
2009	Mathematics	A	25247	441.02	104.60	0.66	439.73	442.31
2009	Mathematics	B	30079	453.54	103.95	0.60	452.36	454.71
2009	Mathematics	C	26798	462.11	99.39	0.61	460.92	463.30
2009	Mathematics	D	32151	479.36	99.33	0.55	478.28	480.45
2009	Mathematics	E	42848	527.55	102.51	0.50	526.58	528.52
2010	Mathematics	A	27710	447.32	107.72	0.65	446.05	448.58
2010	Mathematics	B	32913	460.85	107.40	0.59	459.69	462.01
2010	Mathematics	C	28840	467.07	99.16	0.58	465.93	468.22
2010	Mathematics	D	36217	484.97	100.50	0.53	483.94	486.01
2010	Mathematics	E	48798	531.88	104.43	0.47	530.96	532.81
2011	Mathematics	A	25163	446.30	105.68	0.67	444.99	447.60
2011	Mathematics	B	29934	459.79	106.67	0.62	458.58	461.00
2011	Mathematics	C	26078	462.23	96.26	0.60	461.06	463.40
2011	Mathematics	D	34904	477.63	98.47	0.53	476.59	478.66
2011	Mathematics	E	47746	524.23	104.04	0.48	523.30	525.17
2004	Science	A	8158	448.77	90.82	1.01	446.80	450.74
2004	Science	B	12010	461.41	92.62	0.85	459.75	463.06
2004	Science	C	14197	477.34	92.71	0.78	475.81	478.86
2004	Science	D	9615	492.87	93.72	0.96	491.00	494.74
2004	Science	E	9231	527.04	96.48	1.00	525.07	529.01
2005	Science	A	10731	447.37	105.99	1.02	445.36	449.37
2005	Science	B	12691	455.35	100.75	0.89	453.60	457.10
2005	Science	C	15364	472.58	101.41	0.82	470.98	474.18
2005	Science	D	9533	492.66	104.21	1.07	490.57	494.75
2005	Science	E	7856	531.42	105.29	1.19	529.10	533.75
2006	Science	A	13297	455.03	111.22	0.96	453.14	456.92
2006	Science	B	12564	456.65	103.82	0.93	454.83	458.46
2006	Science	C	15908	474.31	101.60	0.81	472.73	475.89
2006	Science	D	9797	492.45	104.23	1.05	490.39	494.52
2006	Science	E	8351	529.22	104.67	1.15	526.98	531.47
2007	Science	A	14974	444.08	108.34	0.89	442.35	445.82
2007	Science	B	16051	450.29	100.91	0.80	448.73	451.85
2007	Science	C	19115	475.35	100.43	0.73	473.93	476.78
2007	Science	D	10917	502.58	100.97	0.97	500.69	504.48
2007	Science	E	8920	540.60	99.94	1.06	538.52	542.67
2008	Science	A	13799	447.51	112.36	0.96	445.64	449.39
2008	Science	B	16675	449.22	100.82	0.78	447.69	450.75
2008	Science	C	20440	476.78	101.12	0.71	475.40	478.17
2008	Science	D	13165	498.96	100.21	0.87	497.25	500.68
2008	Science	E	10427	539.27	99.14	0.97	537.37	541.18
2009	Science	A	12609	439.78	106.93	0.95	437.91	441.64
2009	Science	B	15116	453.59	103.91	0.85	451.93	455.25

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2009	Science	C	13931	463.62	101.16	0.86	461.94	465.30
2009	Science	D	17191	479.38	100.92	0.77	477.87	480.89
2009	Science	E	25918	519.76	103.04	0.64	518.51	521.02
2010	Science	A	13596	447.87	108.65	0.93	446.04	449.70
2010	Science	B	16449	459.77	107.34	0.84	458.13	461.41
2010	Science	C	14929	466.21	101.27	0.83	464.59	467.84
2010	Science	D	19651	484.11	101.61	0.72	482.69	485.54
2010	Science	E	29778	525.70	101.64	0.59	524.54	526.85
2011	Science	A	11845	445.24	107.02	0.98	443.32	447.17
2011	Science	B	14347	454.89	106.84	0.89	453.14	456.63
2011	Science	C	12923	458.78	99.00	0.87	457.07	460.48
2011	Science	D	18158	473.95	100.05	0.74	472.50	475.41
2011	Science	E	28123	517.34	101.67	0.61	516.15	518.53
2004	History and Social Sciences	A	14109	456.82	94.59	0.80	455.26	458.38
2004	History and Social Sciences	B	19122	467.48	96.30	0.70	466.12	468.85
2004	History and Social Sciences	C	19595	483.08	97.76	0.70	481.71	484.45
2004	History and Social Sciences	D	11919	494.92	102.62	0.94	493.08	496.76
2004	History and Social Sciences	E	10421	528.14	107.00	1.05	526.09	530.20
2005	History and Social Sciences	A	16486	458.91	106.85	0.83	457.28	460.54
2005	History and Social Sciences	B	18876	464.76	102.22	0.74	463.30	466.22
2005	History and Social Sciences	C	20308	480.98	102.00	0.72	479.58	482.38
2005	History and Social Sciences	D	11165	493.58	106.21	1.01	491.61	495.55
2005	History and Social Sciences	E	8691	532.34	107.21	1.15	530.09	534.59
2006	History and Social Sciences	A	19680	458.71	106.68	0.76	457.22	460.20
2006	History and Social Sciences	B	18258	464.50	103.47	0.77	463.00	466.01
2006	History and Social Sciences	C	20329	480.46	102.66	0.72	479.05	481.88
2006	History and Social Sciences	D	11250	496.98	103.28	0.97	495.07	498.89
2006	History and Social Sciences	E	8939	535.09	106.72	1.13	532.87	537.30
2007	History and Social Sciences	A	24591	452.67	105.11	0.67	451.36	453.98
2007	History and Social Sciences	B	25056	463.48	102.77	0.65	462.21	464.75
2007	History and Social Sciences	C	25582	481.14	101.47	0.63	479.89	482.38
2007	History and Social Sciences	D	12262	509.04	101.62	0.92	507.24	510.84

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2007	History and Social Sciences	E	9102	547.38	102.63	1.08	545.27	549.49
2008	History and Social Sciences	A	20335	454.90	108.32	0.76	453.41	456.39
2008	History and Social Sciences	B	24463	458.20	100.19	0.64	456.95	459.46
2008	History and Social Sciences	C	25981	482.78	100.81	0.63	481.55	484.00
2008	History and Social Sciences	D	14443	505.61	101.64	0.85	503.96	507.27
2008	History and Social Sciences	E	10267	543.89	102.58	1.01	541.91	545.88
2009	History and Social Sciences	A	17817	446.34	104.48	0.78	444.80	447.87
2009	History and Social Sciences	B	20941	459.61	103.67	0.72	458.21	461.01
2009	History and Social Sciences	C	18126	468.64	101.35	0.75	467.16	470.11
2009	History and Social Sciences	D	21241	483.89	100.90	0.69	482.54	485.25
2009	History and Social Sciences	E	25874	526.33	103.76	0.65	525.07	527.60
2010	History and Social Sciences	A	19430	452.81	105.30	0.76	451.33	454.29
2010	History and Social Sciences	B	22764	464.92	106.26	0.70	463.54	466.30
2010	History and Social Sciences	C	19473	472.75	100.93	0.72	471.33	474.17
2010	History and Social Sciences	D	23759	488.90	101.67	0.66	487.61	490.19
2010	History and Social Sciences	E	29061	529.35	103.10	0.60	528.16	530.54
2011	History and Social Sciences	A	17480	449.83	103.98	0.79	448.29	451.37
2011	History and Social Sciences	B	20459	461.91	105.30	0.74	460.47	463.35
2011	History and Social Sciences	C	17649	466.26	99.75	0.75	464.79	467.73
2011	History and Social Sciences	D	22819	482.87	100.82	0.67	481.56	484.18
2011	History and Social Sciences	E	28409	522.87	103.09	0.61	521.67	524.07
2004	NEM	A	17586	549.37	98.11	0.74	547.92	550.82
2004	NEM	B	24964	542.94	97.15	0.61	541.74	544.15
2004	NEM	C	26735	542.65	98.05	0.60	541.47	543.82
2004	NEM	D	17264	542.51	97.98	0.75	541.05	543.97
2004	NEM	E	15952	563.65	102.47	0.81	562.06	565.24
2005	NEM	A	21654	552.37	99.73	0.68	551.04	553.70
2005	NEM	B	25240	550.38	96.91	0.61	549.18	551.57
2005	NEM	C	28400	546.10	97.55	0.58	544.97	547.24
2005	NEM	D	16555	545.95	99.22	0.77	544.44	547.46

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
2005	NEM	E	13339	566.22	102.14	0.88	564.49	567.96
2006	NEM	A	26408	547.24	99.16	0.61	546.04	548.44
2006	NEM	B	24648	547.63	97.04	0.62	546.42	548.85
2006	NEM	C	28696	544.27	97.99	0.58	543.14	545.41
2006	NEM	D	16746	544.26	98.45	0.76	542.77	545.75
2006	NEM	E	13670	562.94	101.70	0.87	561.23	564.64
2007	NEM	A	32516	533.99	99.70	0.55	532.90	535.07
2007	NEM	B	34023	533.83	97.34	0.53	532.79	534.86
2007	NEM	C	36589	533.66	97.84	0.51	532.66	534.66
2007	NEM	D	18876	541.29	99.02	0.72	539.88	542.70
2007	NEM	E	14484	560.41	102.54	0.85	558.74	562.08
2008	NEM	A	27394	538.86	100.07	0.60	537.68	540.05
2008	NEM	B	33907	529.42	97.44	0.53	528.38	530.46
2008	NEM	C	37710	530.81	96.95	0.50	529.83	531.79
2008	NEM	D	22427	535.78	98.45	0.66	534.49	537.07
2008	NEM	E	16905	556.09	102.14	0.79	554.55	557.63
2009	NEM	A	24310	528.72	99.84	0.64	527.47	529.98
2009	NEM	B	29531	529.80	99.08	0.58	528.67	530.93
2009	NEM	C	26260	523.89	97.20	0.60	522.71	525.07
2009	NEM	D	31591	524.39	97.37	0.55	523.32	525.46
2009	NEM	E	42078	539.76	100.50	0.49	538.80	540.72
2010	NEM	A	26889	528.89	99.48	0.61	527.70	530.08
2010	NEM	B	32386	529.91	99.70	0.55	528.83	531.00
2010	NEM	C	28312	521.84	96.74	0.57	520.72	522.97
2010	NEM	D	35595	524.36	97.57	0.52	523.35	525.38
2010	NEM	E	47918	539.50	100.07	0.46	538.61	540.40
2011	NEM	A	24696	527.66	100.19	0.64	526.41	528.91
2011	NEM	B	29633	528.62	99.00	0.58	527.49	529.74
2011	NEM	C	25838	518.95	95.37	0.59	517.79	520.11
2011	NEM	D	34621	520.71	96.62	0.52	519.69	521.72
2011	NEM	E	47421	536.66	99.78	0.46	535.77	537.56
2004	LangComm and Mathematics	A	17895	437.35	92.94	0.69	435.99	438.71
2004	LangComm and Mathematics	B	25316	455.35	94.11	0.59	454.19	456.51
2004	LangComm and Mathematics	C	27072	477.35	95.44	0.58	476.21	478.49
2004	LangComm and Mathematics	D	17482	493.31	98.64	0.75	491.85	494.78
2004	LangComm and Mathematics	E	16107	534.77	102.53	0.81	533.19	536.35
2005	LangComm and Mathematics	A	22425	448.15	98.00	0.65	446.87	449.43
2005	LangComm and Mathematics	B	25901	456.75	92.73	0.58	455.62	457.88
2005	LangComm and Mathematics	C	29134	475.53	92.94	0.54	474.46	476.60
2005	LangComm and	D	16947	493.95	97.61	0.75	492.48	495.42

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
	Mathematics							
2005	LangComm and Mathematics	E	13586	537.06	97.70	0.84	535.42	538.70
2006	LangComm and Mathematics	A	27186	453.92	100.20	0.61	452.73	455.11
2006	LangComm and Mathematics	B	25232	457.79	93.79	0.59	456.63	458.95
2006	LangComm and Mathematics	C	29328	475.26	92.71	0.54	474.20	476.32
2006	LangComm and Mathematics	D	17191	494.09	95.39	0.73	492.67	495.52
2006	LangComm and Mathematics	E	14030	537.41	97.73	0.83	535.79	539.03
2007	LangComm and Mathematics	A	33945	447.17	95.49	0.52	446.15	448.18
2007	LangComm and Mathematics	B	35105	456.03	90.28	0.48	455.08	456.97
2007	LangComm and Mathematics	C	37609	477.56	91.26	0.47	476.64	478.48
2007	LangComm and Mathematics	D	19325	506.70	94.19	0.68	505.38	508.03
2007	LangComm and Mathematics	E	14872	549.37	94.38	0.77	547.85	550.88
2008	LangComm and Mathematics	A	28410	449.11	103.27	0.61	447.91	450.32
2008	LangComm and Mathematics	B	34702	451.30	90.24	0.48	450.35	452.25
2008	LangComm and Mathematics	C	38608	478.71	92.27	0.47	477.79	479.63
2008	LangComm and Mathematics	D	22965	503.72	93.01	0.61	502.52	504.93
2008	LangComm and Mathematics	E	17180	547.32	94.13	0.72	545.92	548.73
2009	LangComm and Mathematics	A	25247	439.64	98.38	0.62	438.42	440.85
2009	LangComm and Mathematics	B	30079	454.10	96.18	0.55	453.01	455.19
2009	LangComm and Mathematics	C	26798	463.82	91.61	0.56	462.72	464.92
2009	LangComm and Mathematics	D	32151	480.67	91.93	0.51	479.66	481.67
2009	LangComm and Mathematics	E	42848	527.16	95.31	0.46	526.25	528.06
2010	LangComm and Mathematics	A	27710	445.77	100.50	0.60	444.59	446.96
2010	LangComm and Mathematics	B	32913	460.07	99.68	0.55	458.99	461.15
2010	LangComm and Mathematics	C	28840	467.54	91.97	0.54	466.48	468.61
2010	LangComm and Mathematics	D	36217	485.36	93.24	0.49	484.40	486.32
2010	LangComm and	E	48798	531.39	95.96	0.43	530.53	532.24

Year	PSU Subtest	SES Quintile	N	Mean	S.D.	SEM	Lower C.I.	Upper C.I.
	Mathematics							
2011	LangComm and Mathematics	A	25163	444.33	97.51	0.61	443.12	445.53
2011	LangComm and Mathematics	B	29934	459.19	97.63	0.56	458.09	460.30
2011	LangComm and Mathematics	C	26078	462.99	87.93	0.54	461.92	464.05
2011	LangComm and Mathematics	D	34904	478.85	90.05	0.48	477.91	479.80
2011	LangComm and Mathematics	E	47746	523.89	94.63	0.43	523.04	524.74

Appendix J. Results of Hierarchical Linear Modeling Analysis for Scale Scores

Table 204: Results of Hierarchical Linear Modeling Analysis – Language and Communication All Years

Effect	Region	Estimate	S.E.	DF	t-Value	p
Y_{00}		539.60	1.59	2414	338.71	0.00
NEM		0.11	0.01	231149	19.94	0.00
School SES		2.99	0.11	231149	28.19	0.00
% Female		5.42	2.42	231149	2.24	0.02
Region	Central	18.55	0.92	231149	20.15	0.00
	North	-19.76	1.17	231149	-16.90	0.00
	South	0.00				
School Type	Private	57.33	1.27	231149	45.21	0.00
	Subsidized	19.63	0.96	231149	20.49	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	19.04	0.83	231149	22.92	0.00
	Technical-Professional	0.00				
SES*NEM		0.01	0.00	231149	6.56	0.00
NEM*%*Female		0.06	0.01	231149	8.89	0.00
NEM*Region	Central	-0.05	0.00	231149	-12.83	0.00
	North	-0.02	0.00	231149	-4.15	0.00
	South	0.00				
NEM*Type	Private	0.18	0.01	231149	34.26	0.00
	Subsidized	0.06	0.00	231149	16.39	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.18	0.00	231149	39.85	0.00
	Technical-Professional	0.00				

Table 205: Results of Hierarchical Linear Modeling Analysis – Mathematics All Years

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		552.37	1.72	2414	320.32	0.00
NEM		0.17	0.00	231149	36.05	0.00
School SES		2.75	0.09	231149	28.99	0.00
% Female		-21.25	2.71	231149	-7.83	0.00
Region	Central	5.73	0.92	231149	6.23	0.00
	North	-15.60	1.13	231149	-13.86	0.00
	South	0.00				
School Type	Private	70.31	1.26	231149	55.89	0.00
	Subsidized	18.25	0.94	231149	19.32	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	25.00	0.78	231149	32.22	0.00
	Technical-Professional	0.00				
SES*NEM		0.01	0.00	231149	5.13	0.00
NEM*%*Female		0.00	0.01	231149	-0.87	0.38
NEM*Region	Central	-0.02	0.00	231149	-7.35	0.00
	North	-0.03	0.00	231149	-7.09	0.00
	South	0.00				
NEM*Type	Private	0.13	0.00	231149	28.45	0.00
	Subsidized	0.02	0.00	231149	5.62	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.20	0.00	231149	50.13	0.00
	Technical-Professional	0.00				

Table 206: Results of Hierarchical Linear Modeling Analysis – Science All Years

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		512.38	2.03	2051	251.83	0.00
NEM		0.18	0.01	157521	26.28	0.00
School SES		3.23	0.13	157521	24.77	0.00
% Female		-8.90	3.13	157521	-2.84	0.00
Region	Central	14.69	1.20	157521	12.26	0.00
	North	-12.65	1.48	157521	-8.54	0.00
	South	0.00				
School Type	Private	68.82	1.67	157521	41.28	0.00
	Subsidized	18.86	1.25	157521	15.08	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	29.26	1.15	157521	25.47	0.00
	Technical-Professional	0.00				
SES*NEM		0.01	0.00	157521	4.13	0.00
NEM*%*Female		0.02	0.01	157521	2.34	0.02
NEM*Region	Central	-0.03	0.00	157521	-7.57	0.00
	North	-0.05	0.01	157521	-8.92	0.00
	South	0.00				
NEM*Type	Private	0.14	0.01	157521	21.41	0.00
	Subsidized	0.02	0.00	157521	5.67	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.26	0.01	157521	41.59	0.00
	Technical-Professional	0.00				

Table 207: Results of Hierarchical Linear Modeling Analysis – History and Social Sciences All Years

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		556.98	1.99	2050	279.24	0.00
NEM		0.03	0.01	137708	3.65	0.00
School SES		2.21	0.17	137708	13.36	0.00
% Female		-17.60	2.84	137708	-6.19	0.00
Region	Central	22.14	1.22	137708	18.21	0.00
	North	-21.56	1.59	137708	-13.53	0.00
	South	0.00				
School Type	Private	49.92	1.69	137708	29.57	0.00
	Subsidized	13.88	1.27	137708	10.95	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	13.83	1.15	137708	12.07	0.00
	Technical-Professional	0.00				
SES*NEM		0.01	0.00	137708	3.66	0.00
NEM*%*Female		0.09	0.01	137708	8.34	0.00
NEM*Region	Central	-0.01	0.01	137708	-2.33	0.02
	North	0.00	0.01	137708	0.05	0.96
	South	0.00				
NEM*Type	Private	0.22	0.01	137708	26.48	0.00
	Subsidized	0.06	0.01	137708	10.07	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.12	0.01	137708	18.44	0.00
	Technical-Professional	0.00				

Appendix K. Factorial Analysis of Variance of PSU Subtest Raw Scores by Year, Gender, Type, Region and Curricular Branch

Table 208: Factorial Analysis of Variance—Language and Communication & Mathematics

PSU Test	Source	DF	SS	F	p	Effect Size (f)
LangComm & Mathematics	Year	8	8970019	1223.81	0.00	0.08
LangComm & Mathematics	Gender	1	8141189	8885.86	0.00	0.08
LangComm & Mathematics	Year*Gender	8	41097	5.61	0.00	0.01
LangComm & Mathematics	Error	1457280	1335154712			
LangComm & Mathematics	Year	8	9063443	1237.54	0.00	0.08
LangComm & Mathematics	Region	2	9419092	5144.41	0.00	0.08
LangComm & Mathematics	Year*Region	16	604104	41.24	0.00	0.02
LangComm & Mathematics	Error	1452845	1330033366			
LangComm & Mathematics	Year	8	8982990	1480.15	0.00	0.09
LangComm & Mathematics	Type	3	236691666	104000.76	0.00	0.46
LangComm & Mathematics	Year*Type	16	2602408	214.40	0.00	0.05
LangComm & Mathematics	Error	1443234	1094868498			
LangComm & Mathematics	Year	8	8969602	1461.68	0.00	0.09
LangComm & Mathematics	Branch	8	224607993	36602.01	0.00	0.45
LangComm & Mathematics	Year*Branch	56	942096	21.93	0.00	0.03
LangComm & Mathematics	Error	1457211	1117770844			
LangComm & Mathematics	Year	8	7767014	1389.05	0.00	0.09
LangComm & Mathematics	SES Quintile	4	72047553	25769.92	0.00	0.29
LangComm & Mathematics	Year*SES	32	9387279	419.70	0.00	0.10
LangComm & Mathematics	Error	1247151	871696343			

Table 209: Factorial Analysis of Variance—Language and Communication

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Language and Communication	Year	8	2804319	1494.56	0.00	0.09
Language and Communication	Gender	1	72618	309.61	0.00	0.01
Language and Communication	Year*Gender	8	71696	38.21	0.00	0.01
Language and Communication	Error	1457185	341774295			
Language and Communication	Year	8	2824892	1517.28	0.00	0.09
Language and Communication	Region	2	2858871	6142.13	0.00	0.09
Language and Communication	Year*Region	16	95657	25.69	0.00	0.02
Language and Communication	Error	1452750	338093272			
Language and Communication	Year	8	2794927	1720.93	0.00	0.10
Language and Communication	Type	3	45433569	74600.12	0.00	0.39
Language and Communication	Year*Type	16	823533	253.54	0.00	0.05
Language and Communication	Error	1443139	292970753			
Language and Communication	Year	8	2803882	1735.38	0.00	0.10
Language and Communication	Branch	8	47029632	29107.65	0.00	0.40
Language and Communication	Year*Branch	56	599394	53.00	0.00	0.04
Language and Communication	Error	1457116	294285325			
Language and Communication	Year	8	2355726	1524.49	0.00	0.10
Language and Communication	SES Quintile	4	17194314	22254.27	0.00	0.27
Language and Communication	Year*SES	32	1995364	322.82	0.00	0.09
Language and Communication	Error	1247062	240879344			

Table 210: Factorial Analysis of Variance—Mathematics

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Mathematics	Year	8	2966404	1295.67	0.00	0.08
Mathematics	Gender	1	6667530	23298.02	0.00	0.13
Mathematics	Year*Gender	8	22268	9.73	0.00	0.01
Mathematics	Error	1454752	416327271			
Mathematics	Year	8	2990920	1291.39	0.00	0.08
Mathematics	Region	2	1907548	3294.50	0.00	0.07
Mathematics	Year*Region	16	222616	48.06	0.00	0.02
Mathematics	Error	1450325	419876125			
Mathematics	Year	8	2970814	1550.90	0.00	0.09
Mathematics	Type	3	74719537	104018.84	0.00	0.47
Mathematics	Year*Type	16	543526	141.87	0.00	0.04
Mathematics	Error	1440744	344975088			
Mathematics	Year	8	2966538	1512.10	0.00	0.09
Mathematics	Branch	8	66209034	33748.04	0.00	0.43
Mathematics	Year*Branch	56	66924	4.87	0.00	0.01
Mathematics	Error	1454683	356736106			
Mathematics	Year	8	2569755	1482.25	0.00	0.10
Mathematics	SES Quintile	4	18811709	21701.41	0.00	0.26
Mathematics	Year*SES	32	2806132	404.65	0.00	0.10
Mathematics	Error	1244718	269743725			

Table 211: Factorial Analysis of Variance—Science

PSU Test	Source	DF	SS	F	p	Effect Size (f)
Science	Year	8	1752376	833.44	0.00	0.09
Science	Gender	1	2421767	9214.49	0.00	0.11
Science	Year*Gender	8	78930	37.54	0.00	0.02
Science	Error	786159	206619512			
Science	Year	8	1778269	849.93	0.00	0.09
Science	Region	2	3439721	6576.13	0.00	0.13
Science	Year*Region	16	160541	38.37	0.00	0.03
Science	Error	784145	205078132			
Science	Year	8	1741112	957.28	0.00	0.10
Science	Type	3	30075497	44095.39	0.00	0.41
Science	Year*Type	16	634767	174.50	0.00	0.06
Science	Error	779379	177193215			
Science	Year	8	1752006	929.64	0.00	0.10
Science	Branch	8	23870853	12666.24	0.00	0.36
Science	Year*Branch	56	62839	4.76	0.00	0.02
Science	Error	786093	185184358			
Science	Year	8	1476978	909.85	0.00	0.11
Science	SES Quintile	4	6500420	8008.75	0.00	0.22
Science	Year*SES	32	1848493	284.68	0.00	0.12
Science	Error	655869	133086381			

Table 212: Factorial Analysis of Variance—History and Social Sciences

PSU Test	Source	DF	SS	F	p	Effect Size (f)
History	Year	8	1120832	572.61	0.00	0.07
History	Gender	1	3038186	12417.07	0.00	0.11
History	Year*Gender	8	91218	46.60	0.00	0.02
History	Error	965447	236223797			
History	Year	8	1157020	590.06	0.00	0.07
History	Region	2	2736538	5582.36	0.00	0.11
History	Year*Region	16	147216	37.54	0.00	0.02
History	Error	962507	235915970			
History	Year	8	1146841	655.87	0.00	0.07
History	Type	3	27616460	42116.50	0.00	0.36
History	Year*Type	16	834955	238.75	0.00	0.06
History	Error	955976	208949545			
History	Year	8	1120692	628.89	0.00	0.07
History	Branch	8	23535169	13206.97	0.00	0.33
History	Year*Branch	56	773640	62.02	0.00	0.06
History	Error	965384	215042544			
History	Year	8	1157864	696.61	0.00	0.08
History	SES Quintile	4	8610253	10360.41	0.00	0.22
History	Year*SES	32	1864851	280.49	0.00	0.10
History	Error	838285	174168935			

Appendix L. Trend Analysis of the PSU Raw Scores by Subtest and by Subpopulation

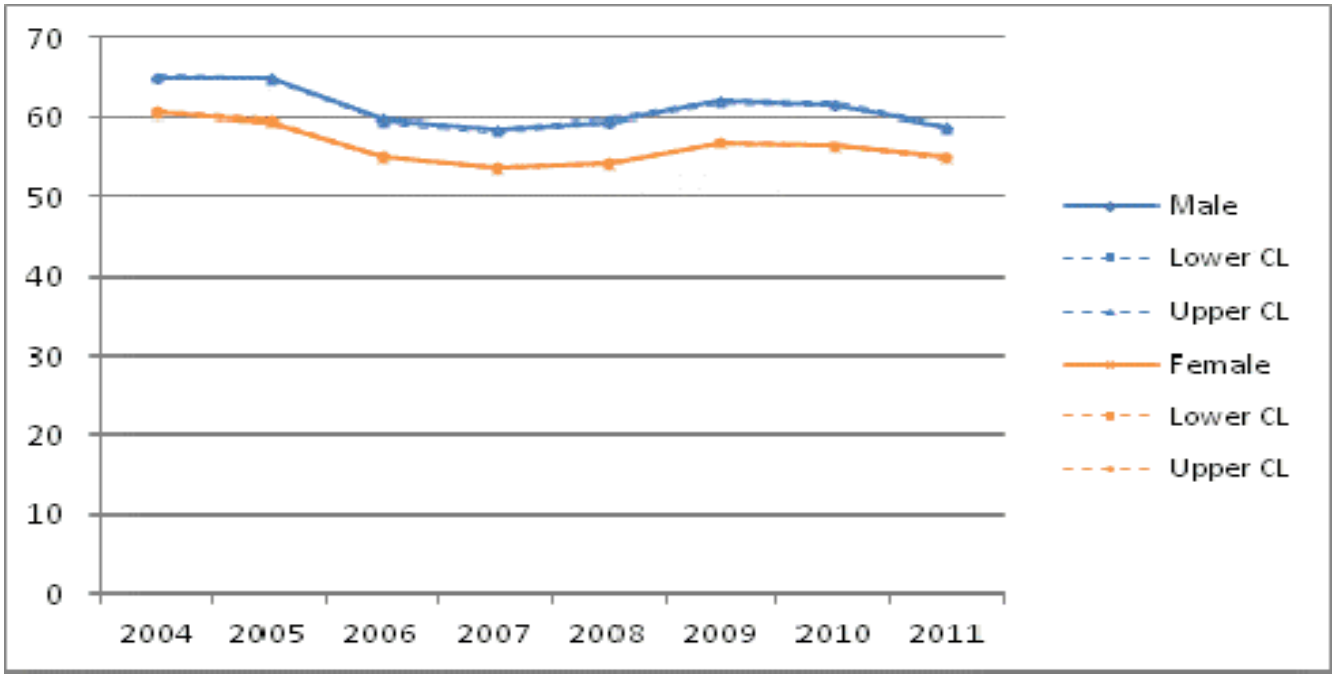


Figure 81: Language and Communication and Mathematics Combined Raw Score by Gender

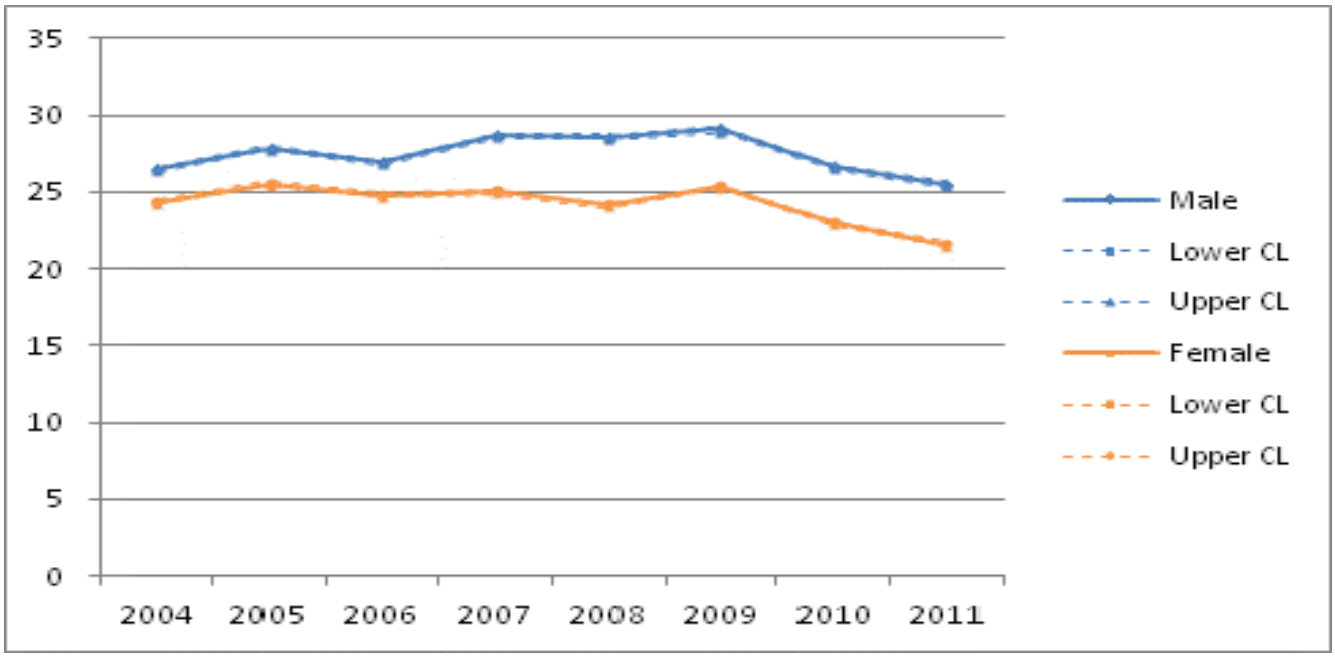


Figure 82: Science Raw Score by Gender

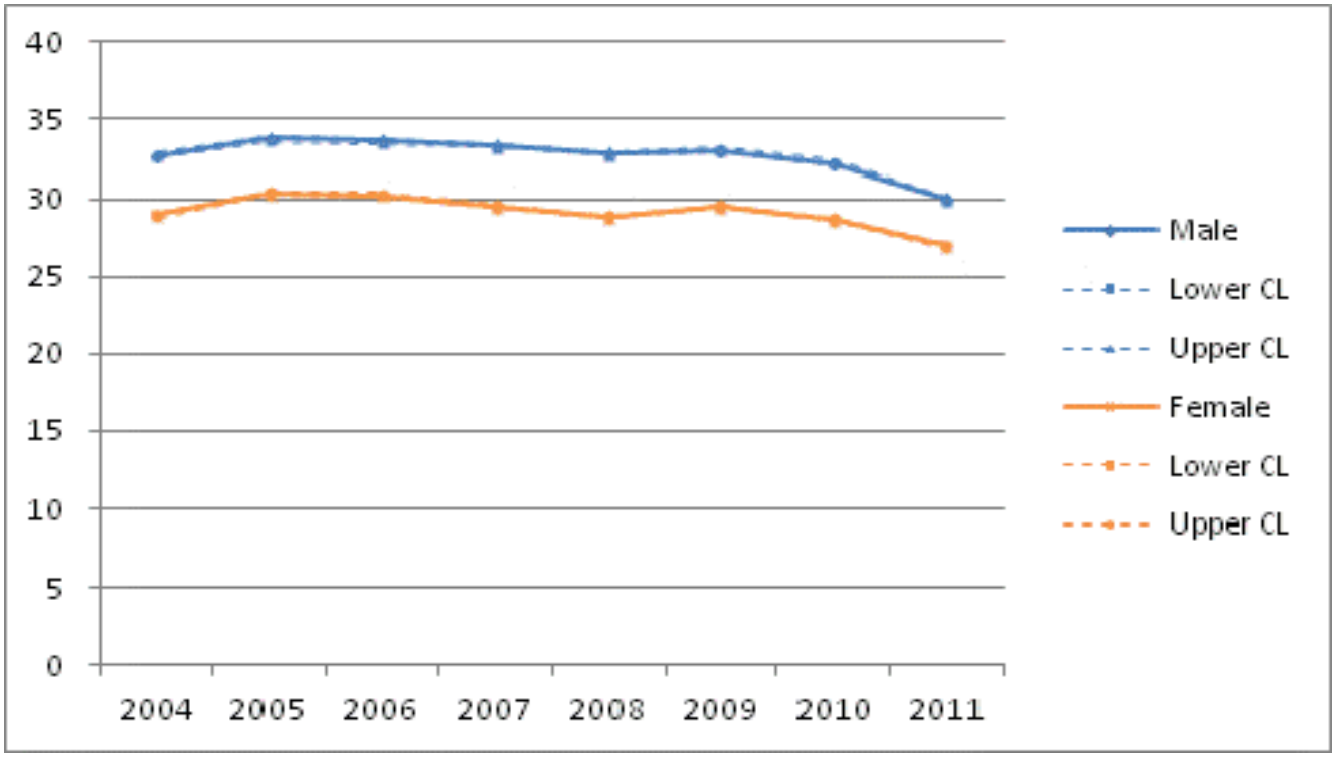


Figure 83: History and Social Sciences Raw Score by Gender

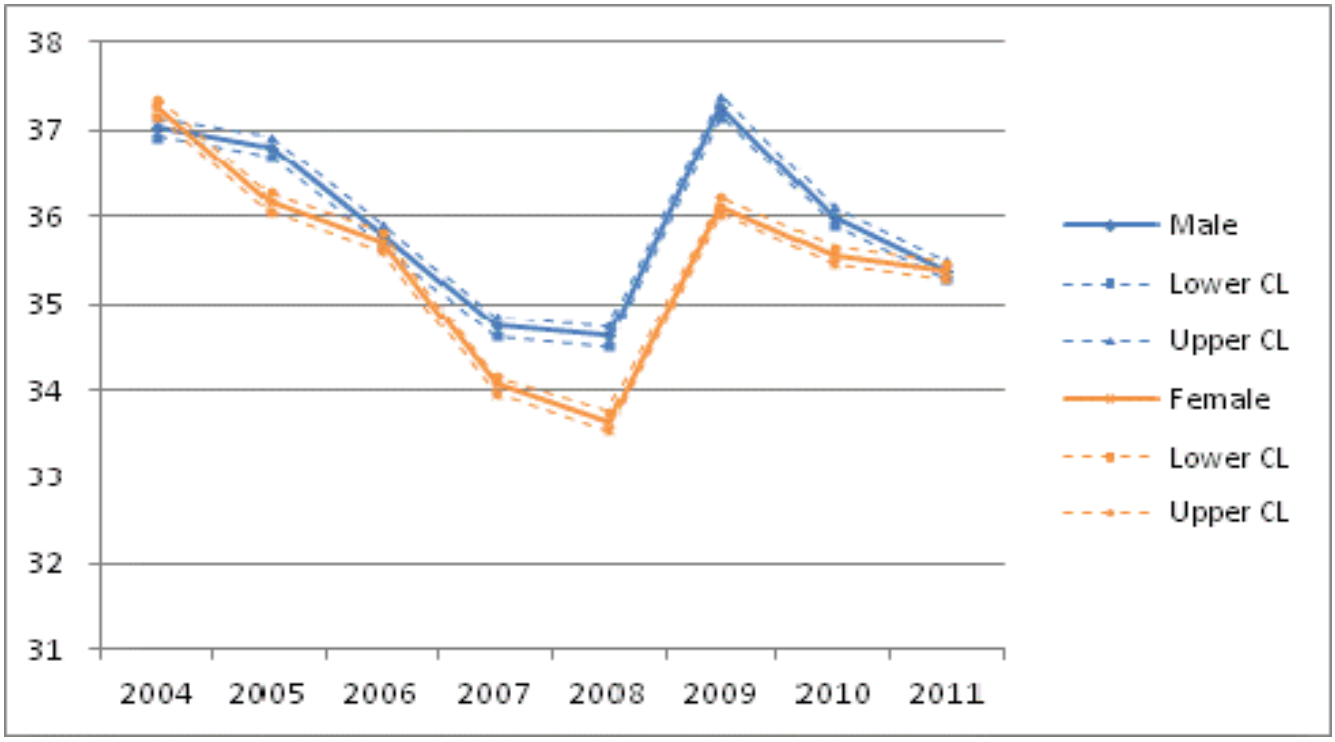


Figure 84: Language and Communication Raw Score by Gender

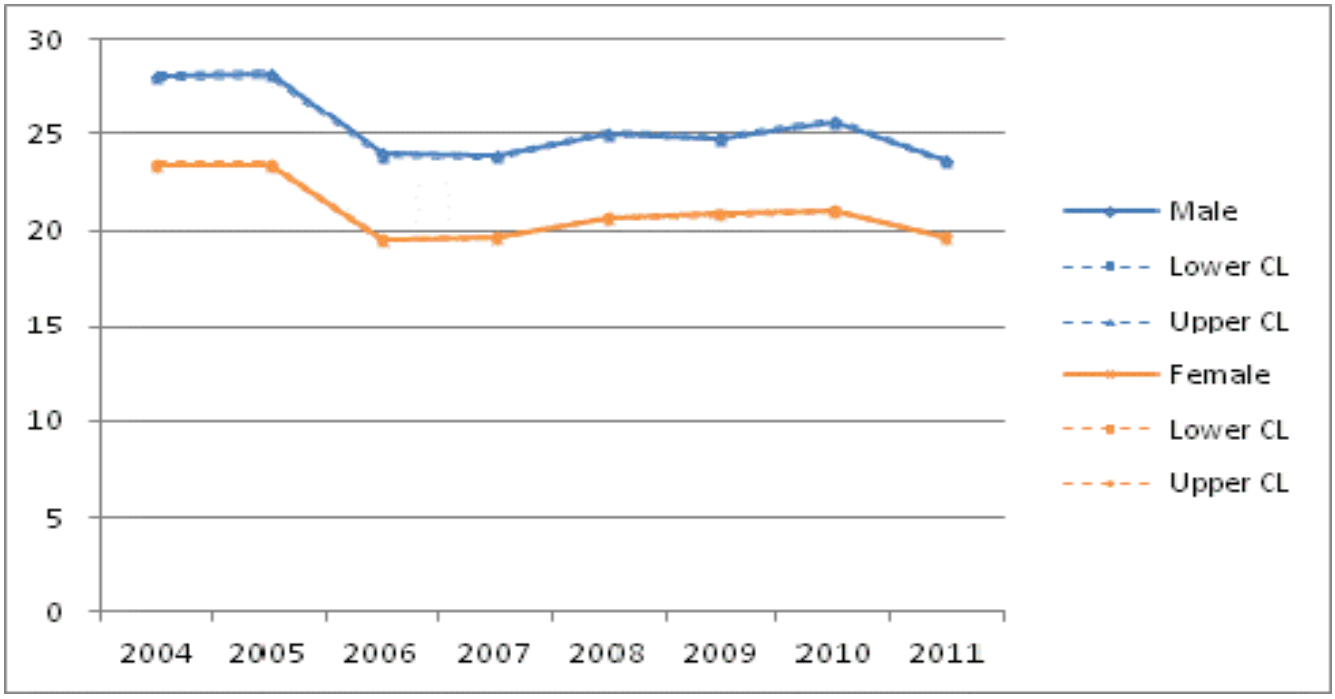


Figure 85: Mathematics Raw Score by Gender

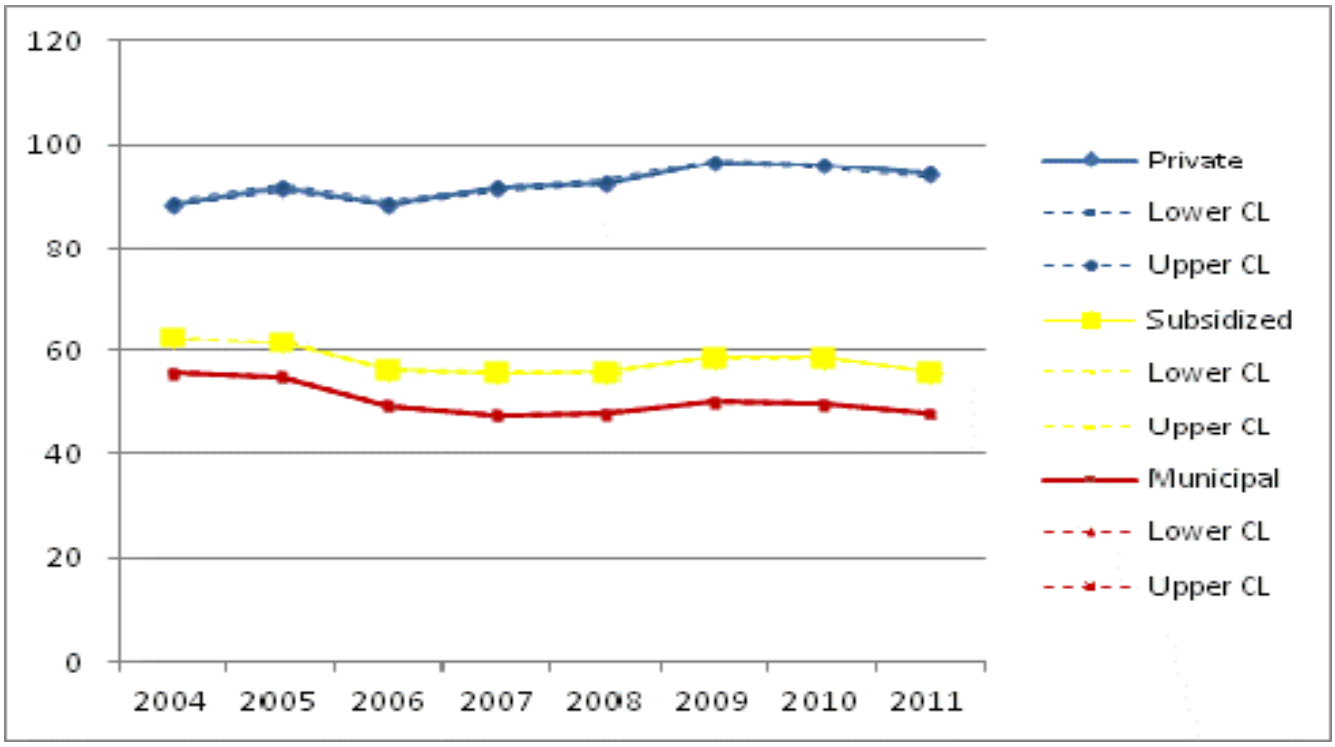


Figure 86: Language and Communication and Mathematics Raw Score by School Type

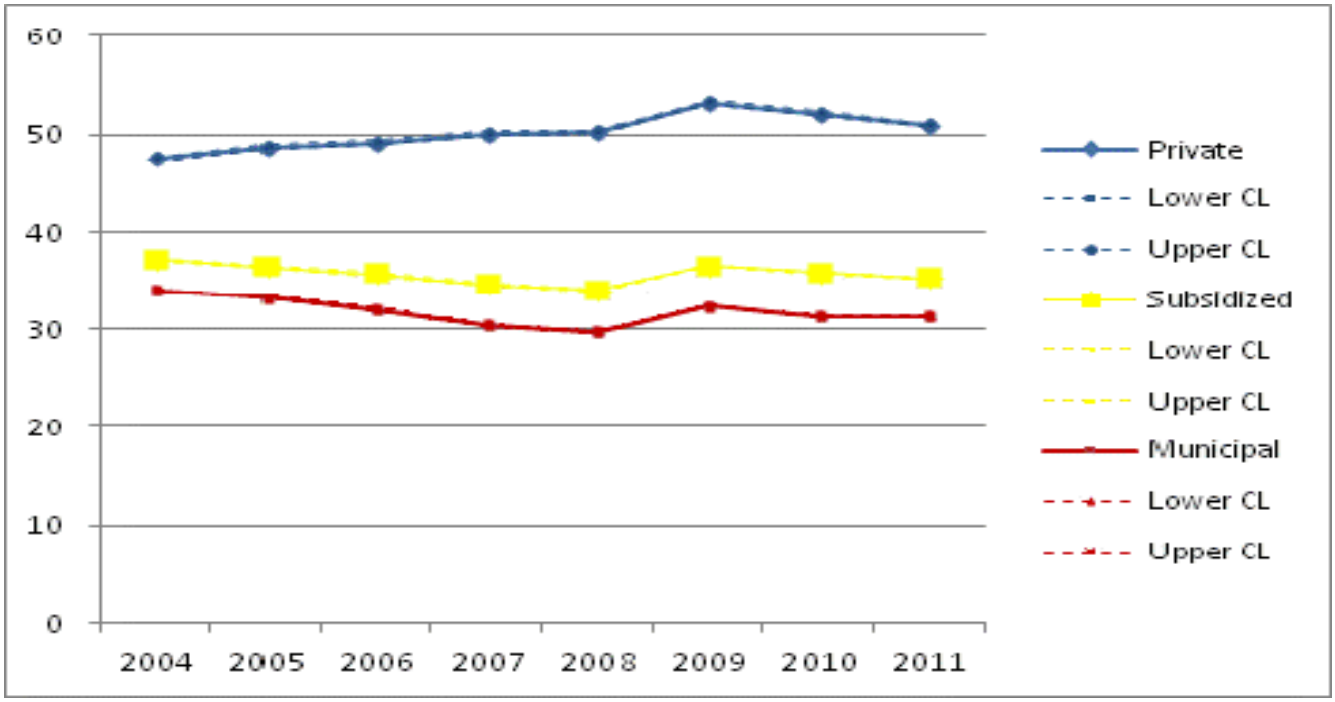


Figure 87: Language and Communication Raw Score by School Type

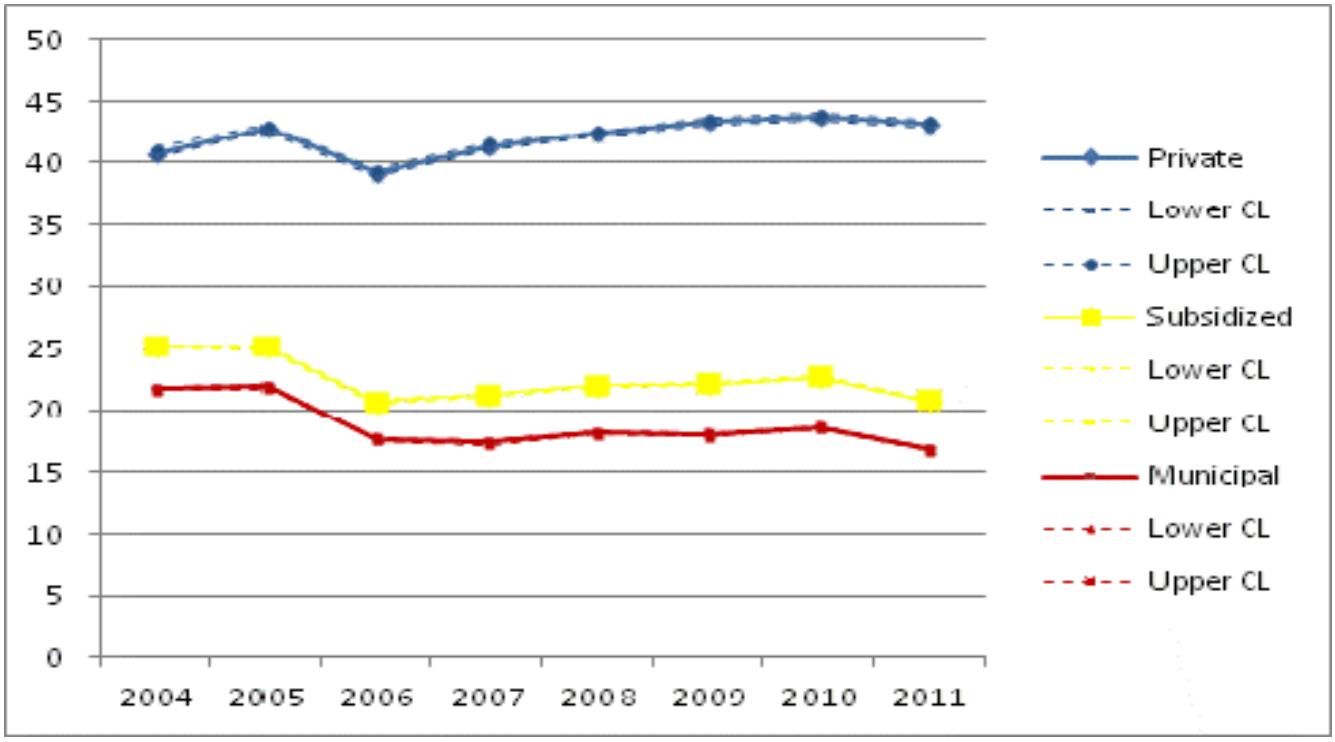


Figure 88: Mathematics Raw Score by School Type

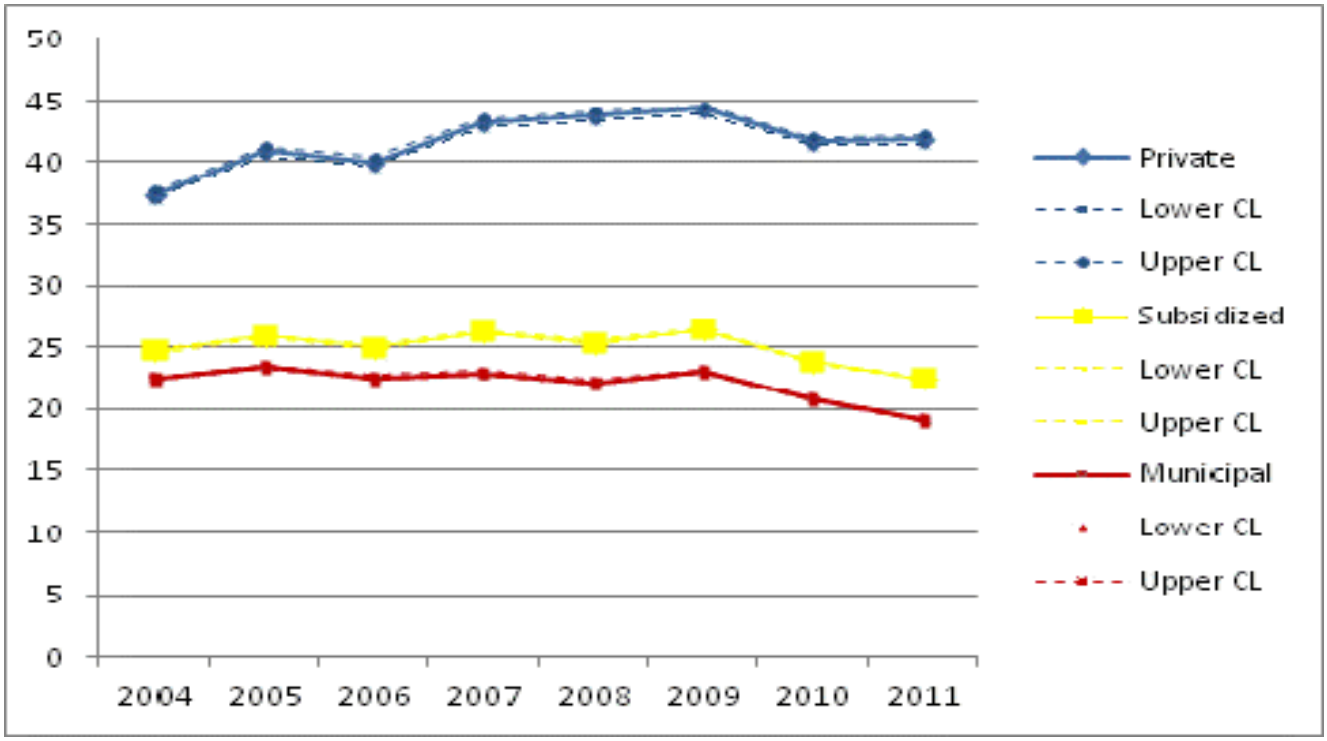


Figure 89: Science Raw Score by School Type

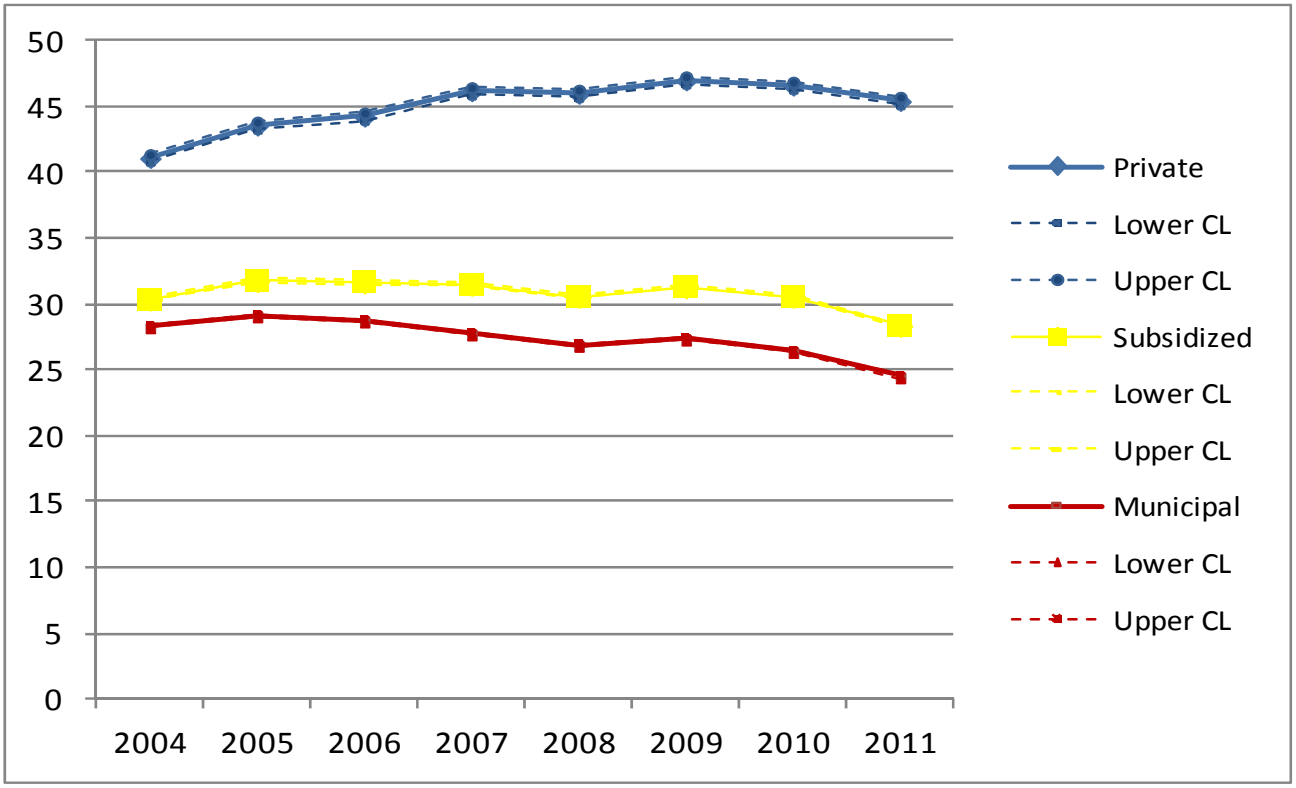


Figure 90: History and Social Sciences Raw Score by School Type

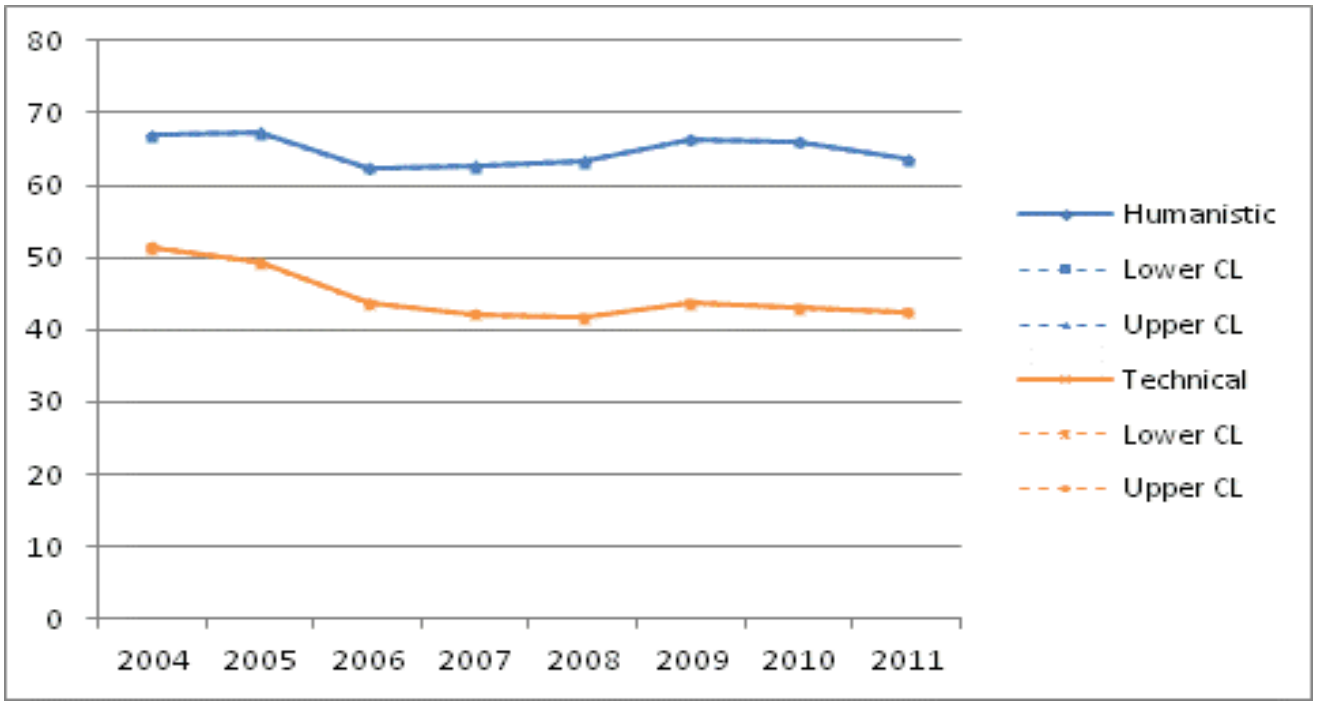


Figure 91: Language and Communication and Mathematics Raw Score by Curricular Branch

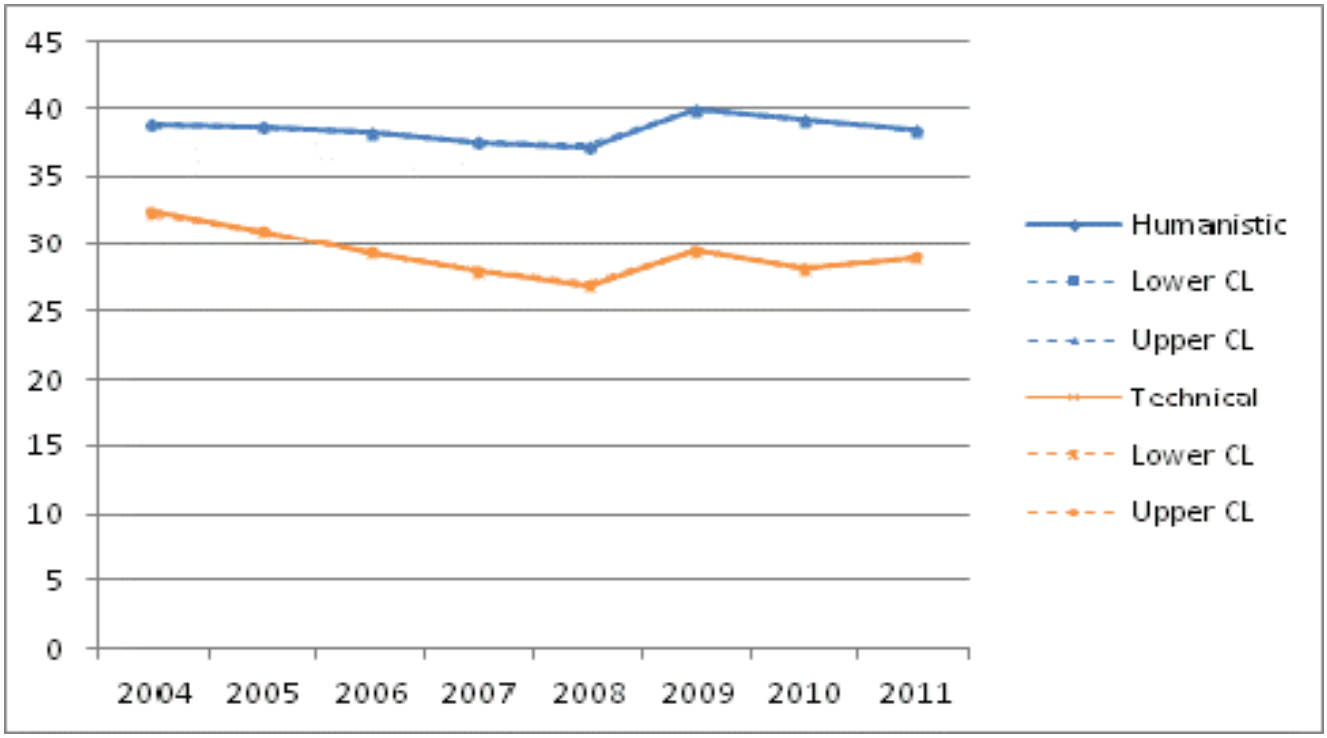


Figure 92: Language and Communication Raw Score by Curricular Branch

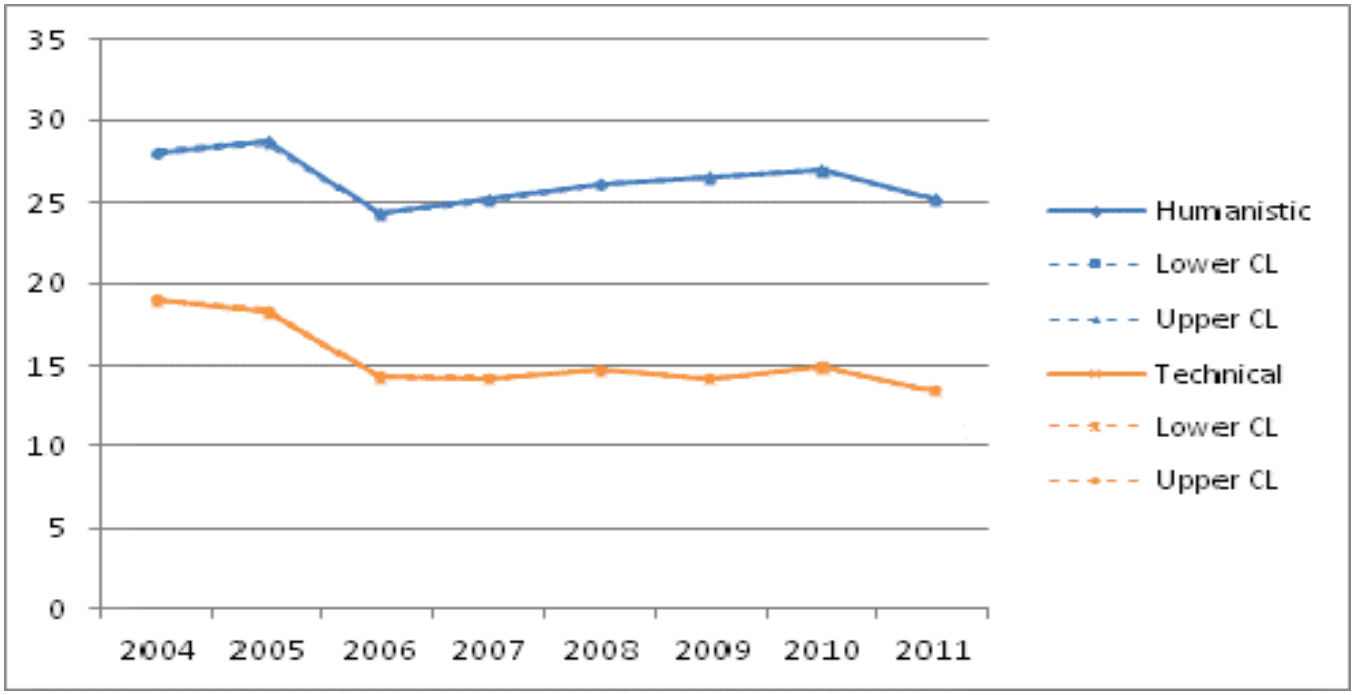


Figure 93: Mathematics Raw Score by Curricular Branch

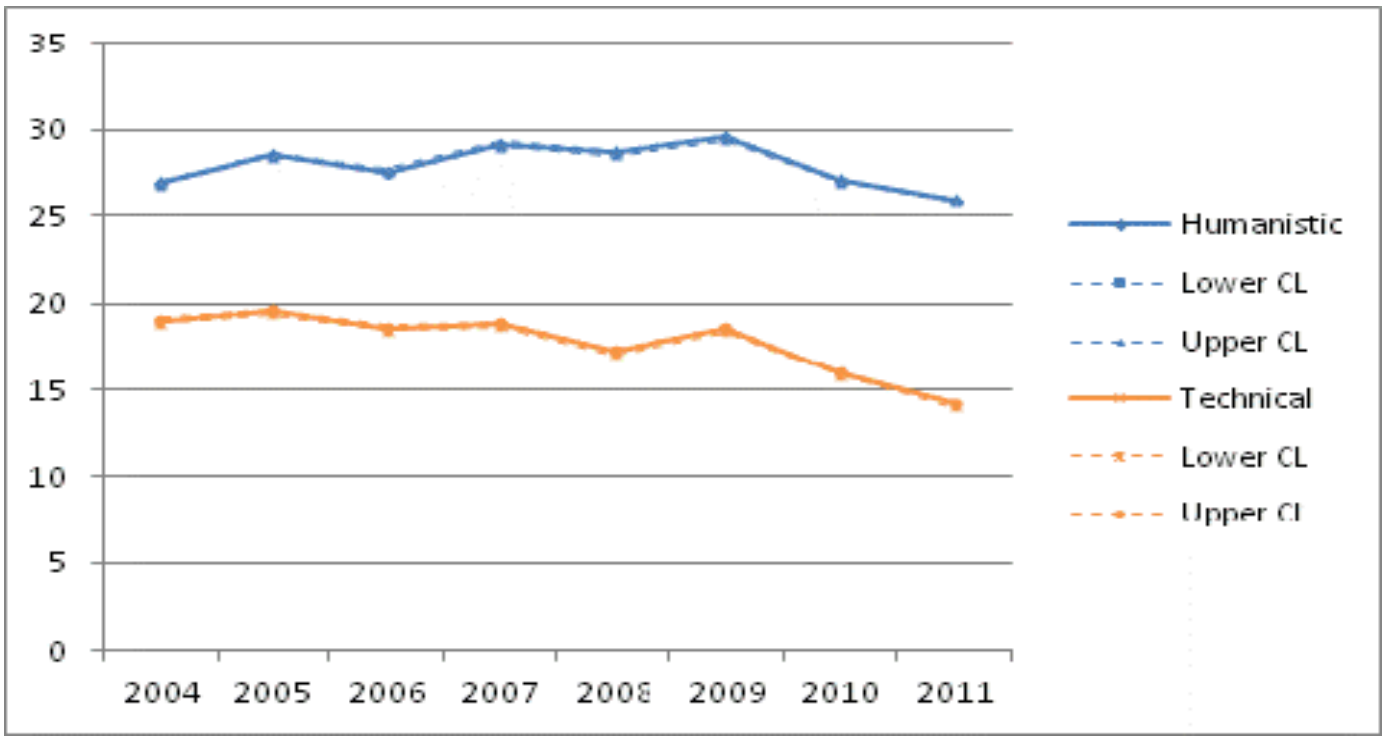


Figure 94: Science Raw Score by Curricular Branch

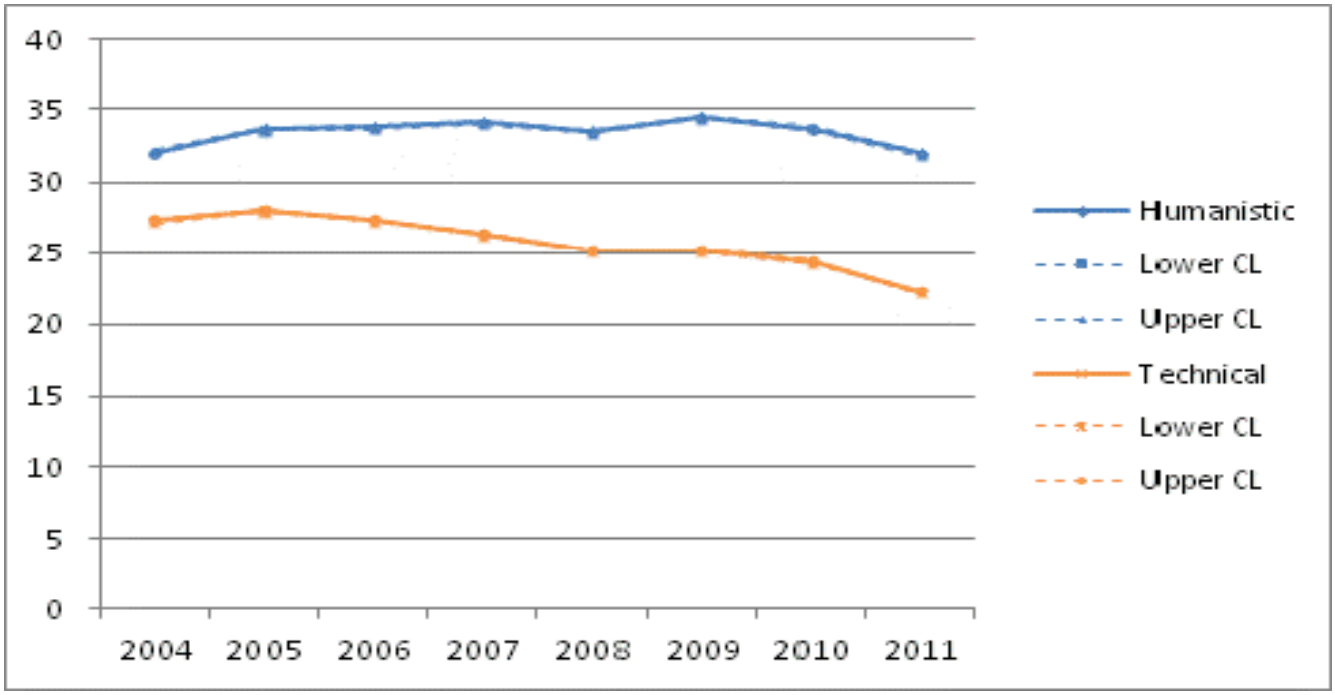


Figure 95: History and Social Sciences Raw Score by Curricular Branch

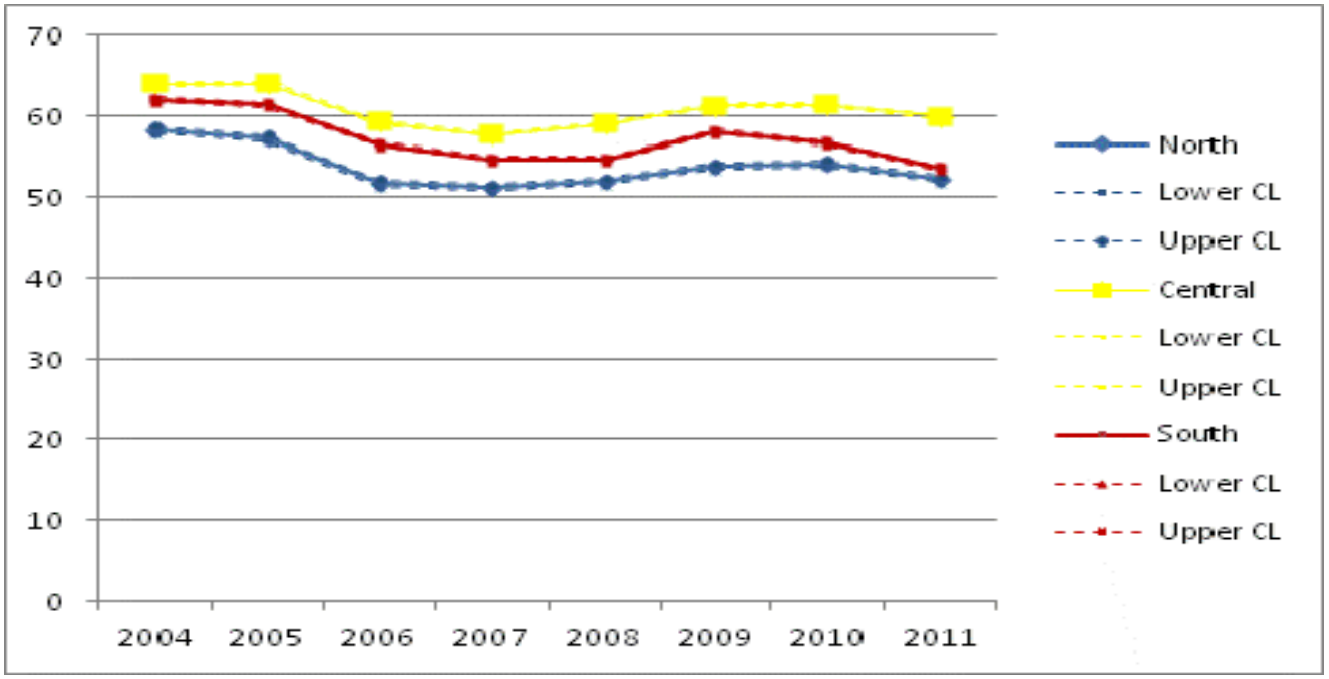


Figure 96: Language and Communication and Mathematics Raw Score by Region

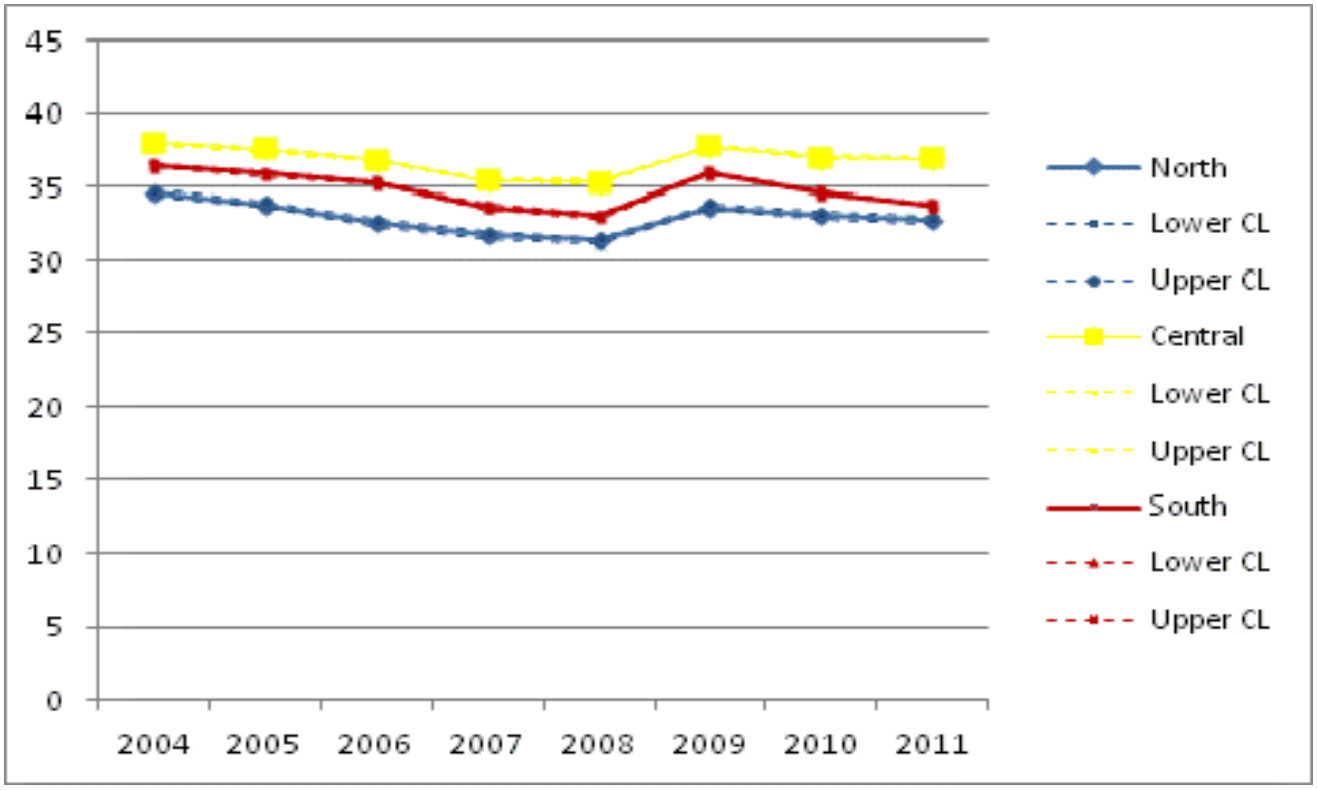


Figure 97: Language and Communication Raw Score by Region

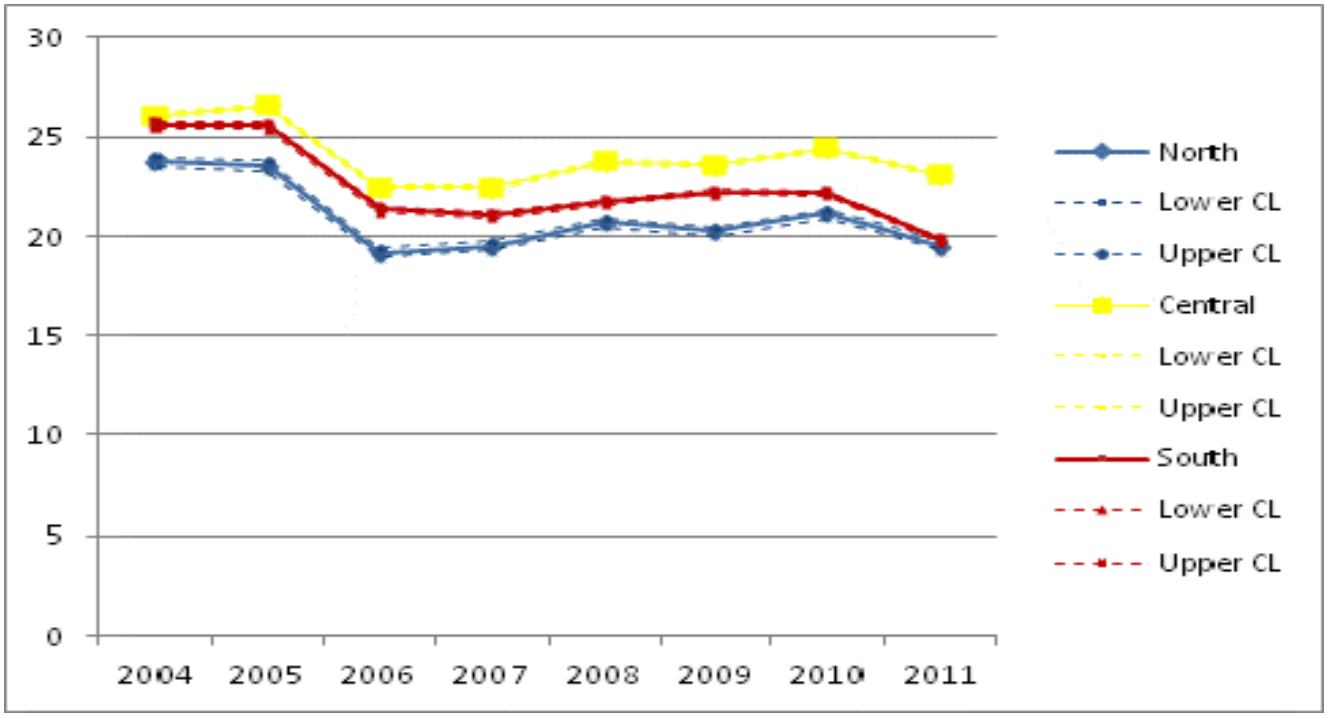


Figure 98: Mathematics Raw Score by Region

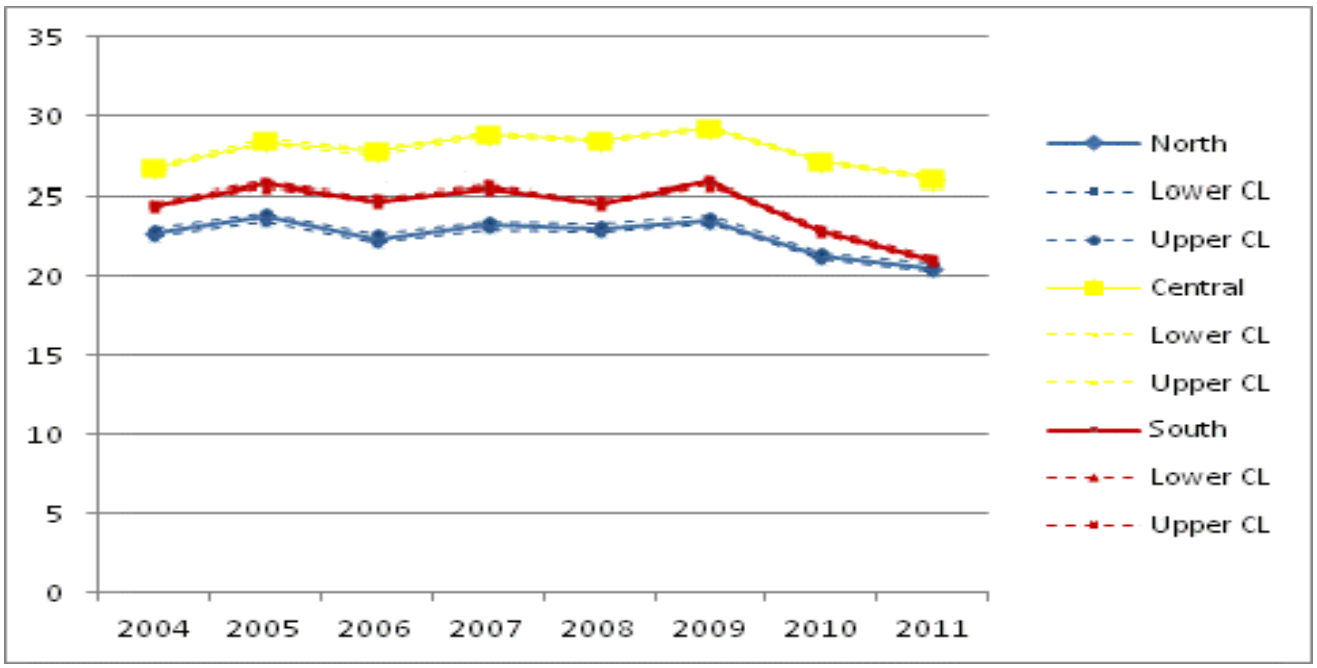


Figure 99: Science Raw Score by Region

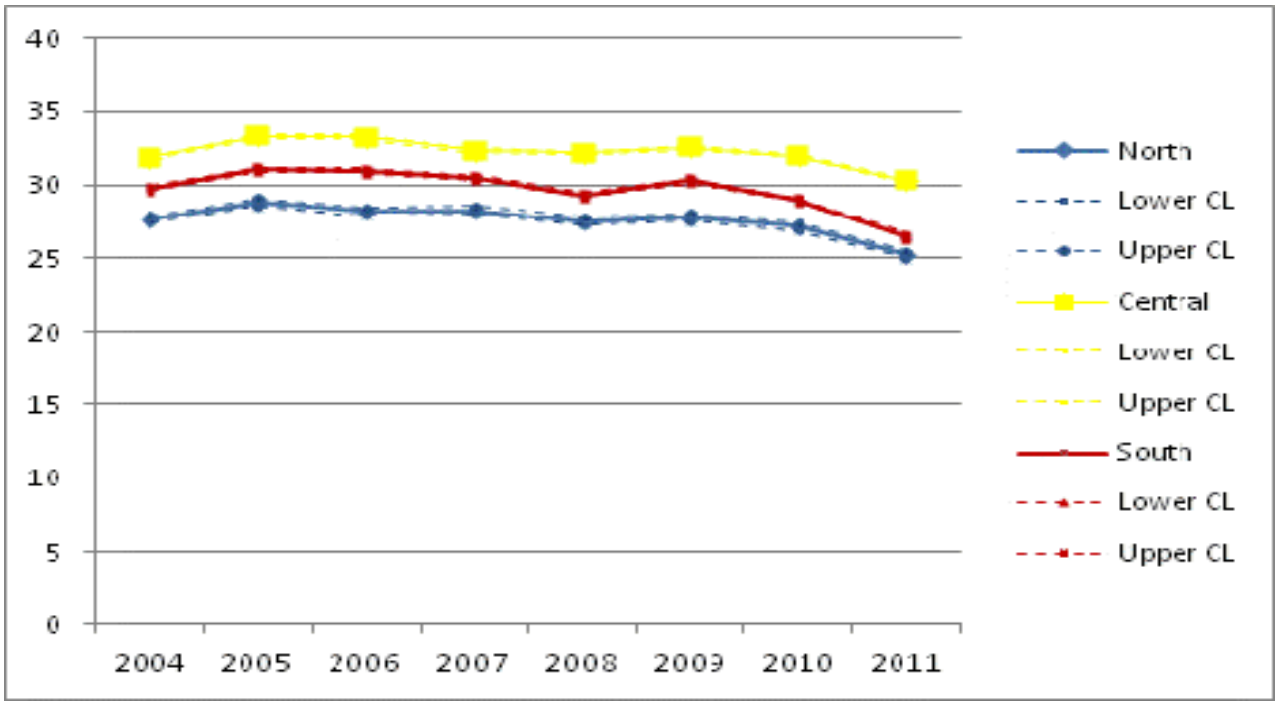


Figure 100: History and Social Sciences Raw Score by Region

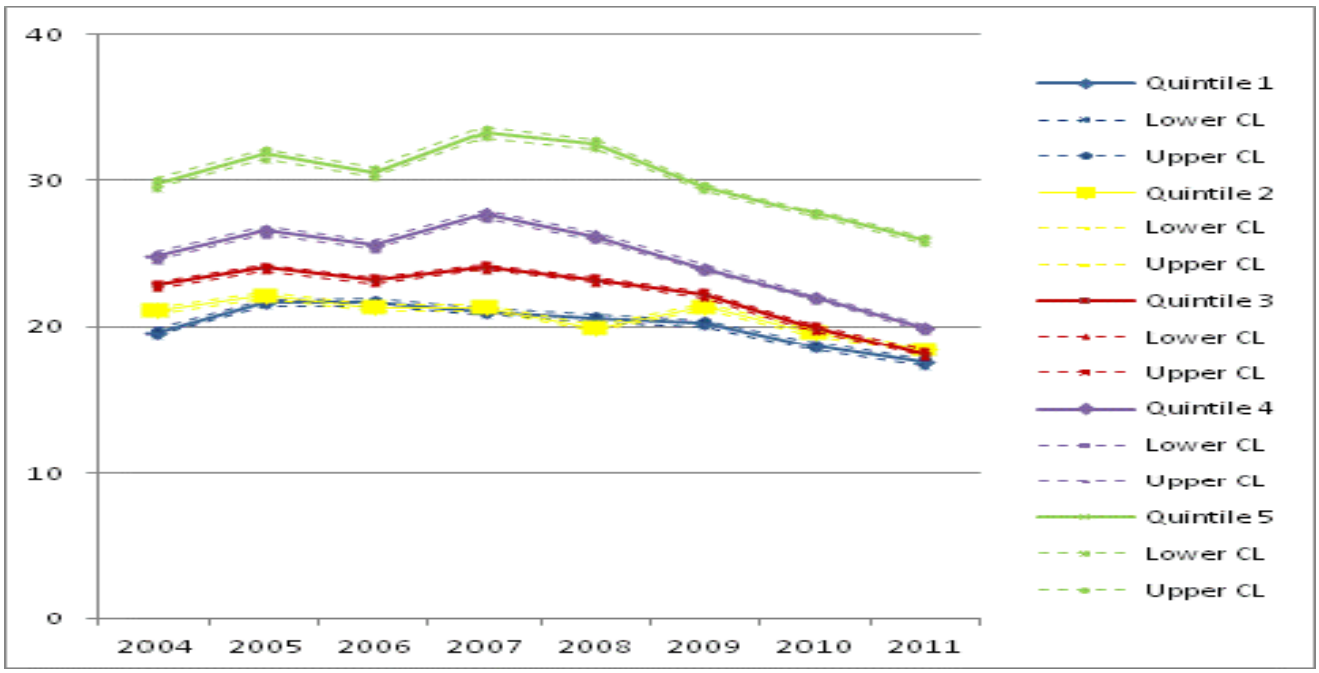


Figure 101: Science Raw Score by Region

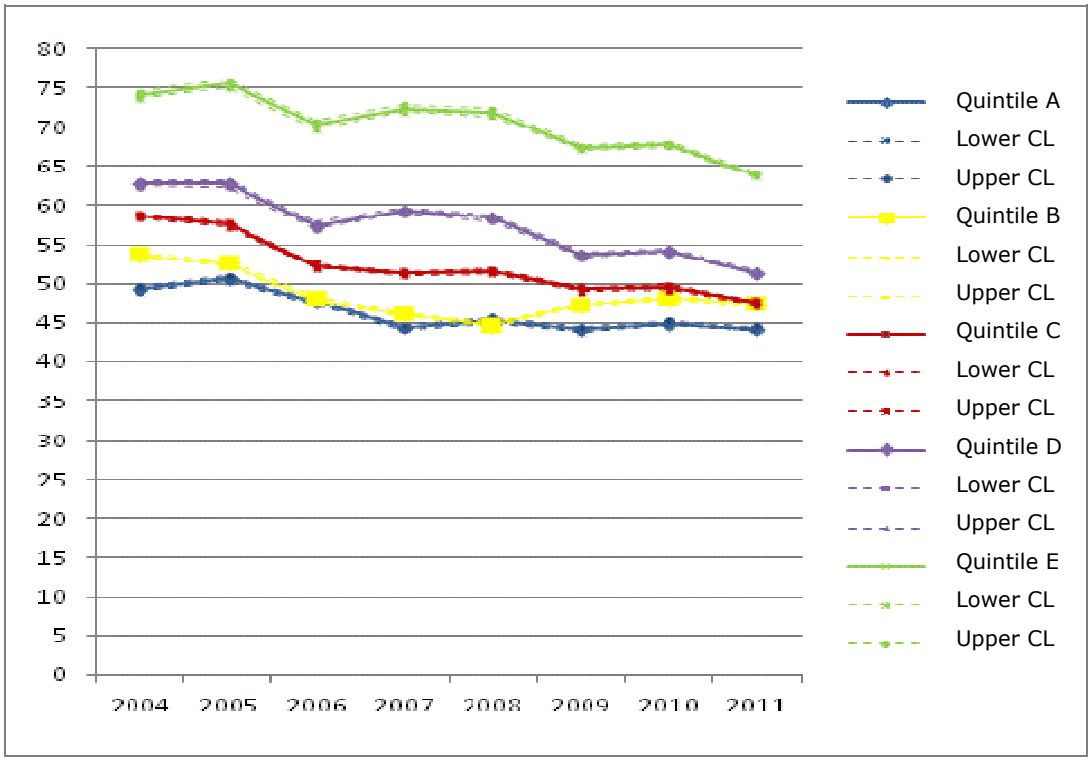


Figure 102: Language and Communication and Mathematics Raw Score by SES Quintile

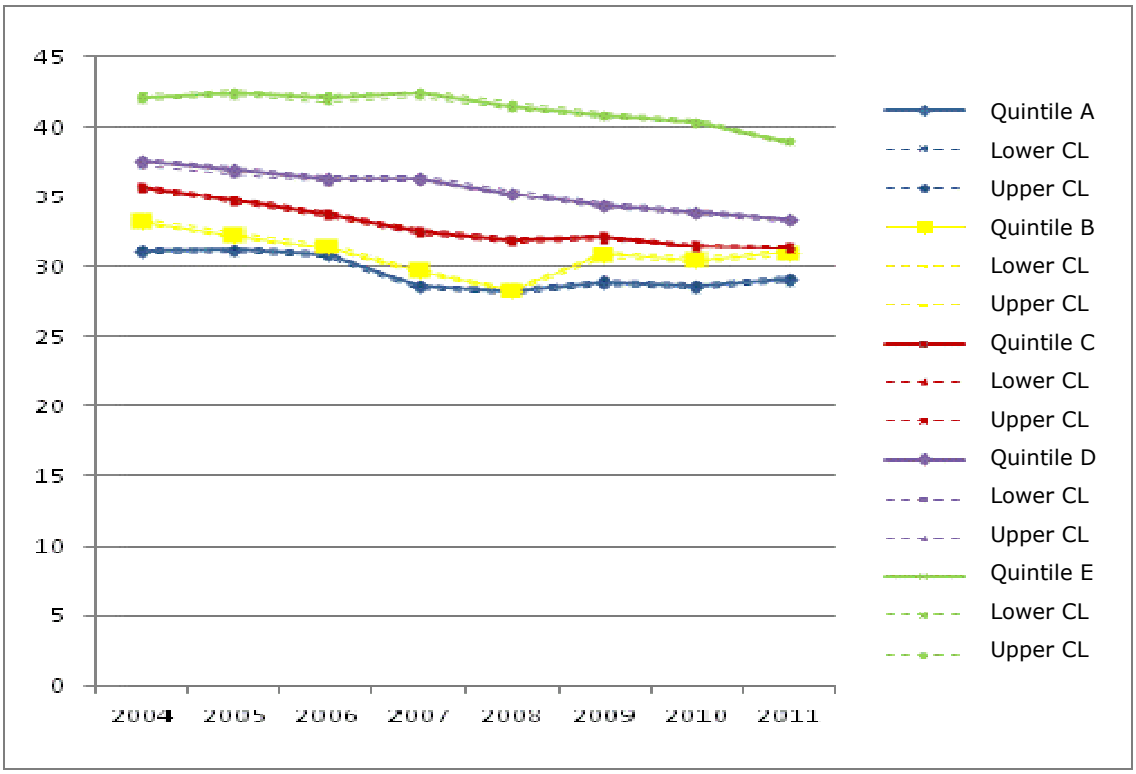


Figure 103: Language and Communication Raw Score by SES Quintile

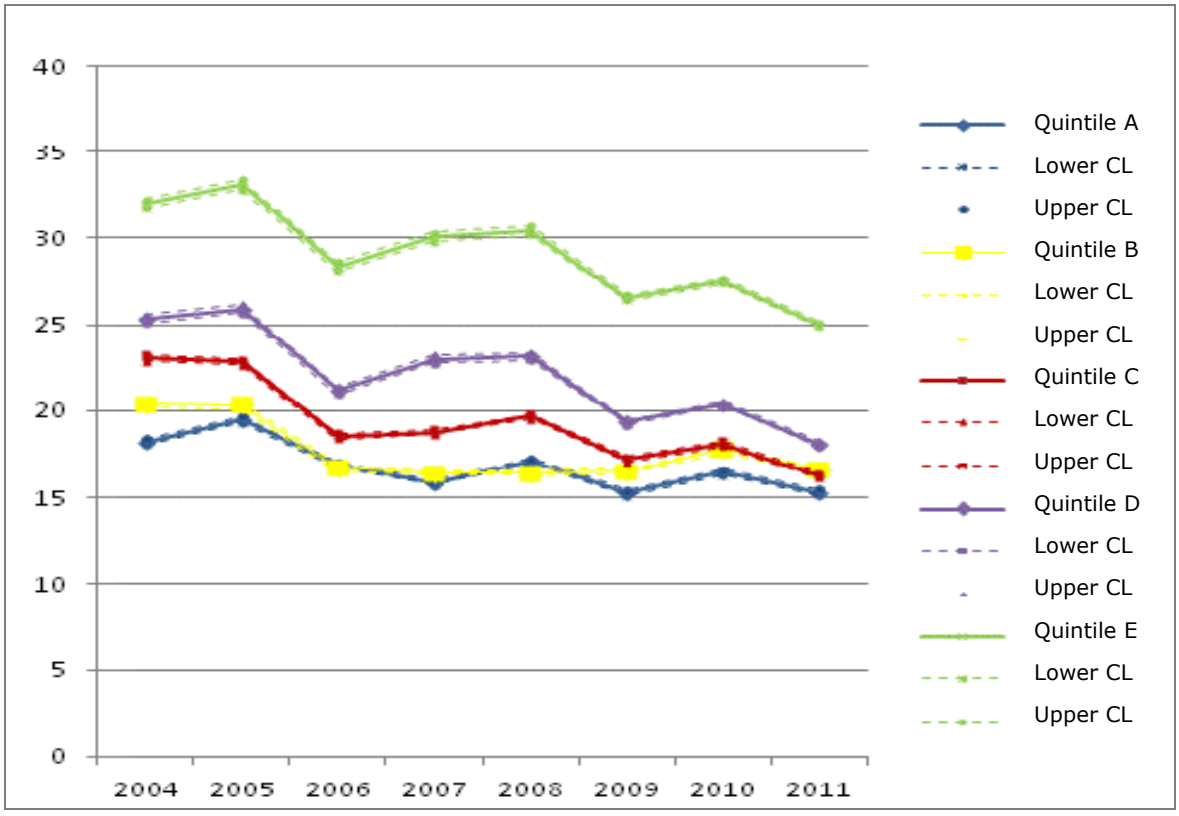


Figure 104: Mathematics Raw Score by SES Quintile

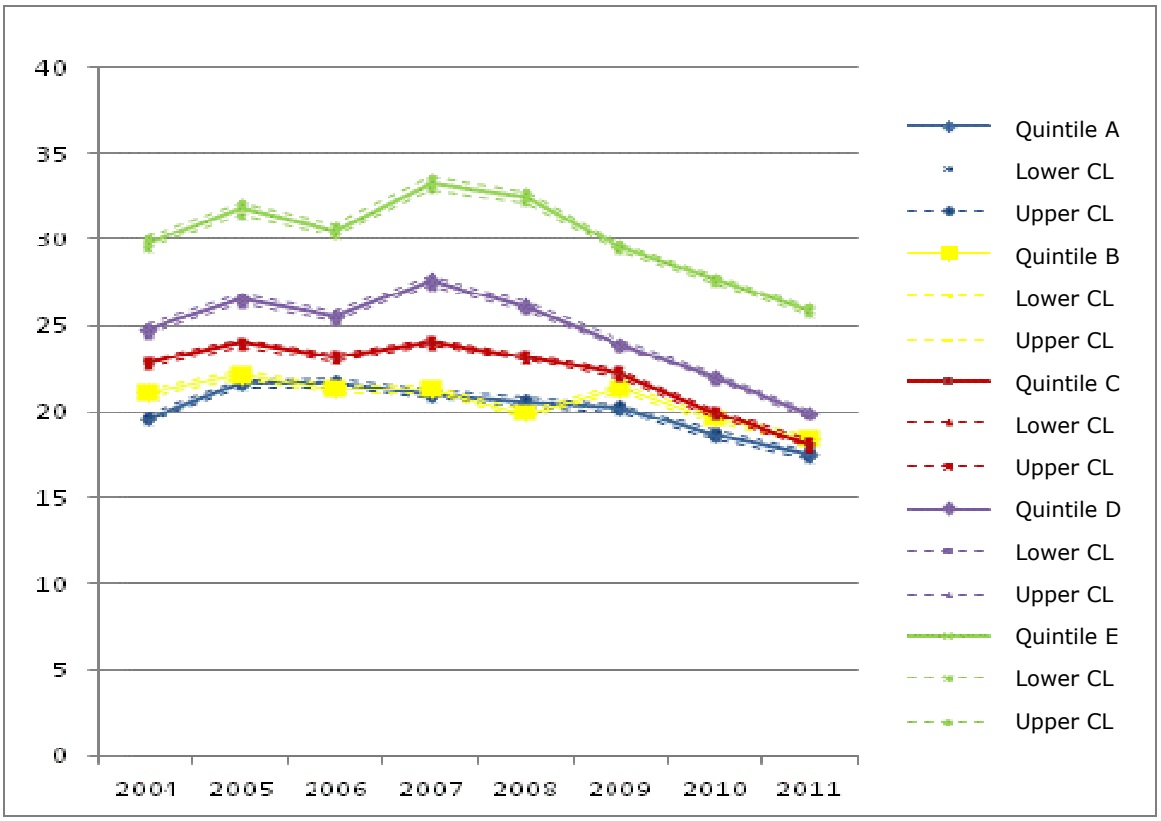


Figure 105: Science Raw Score by SES Quintile

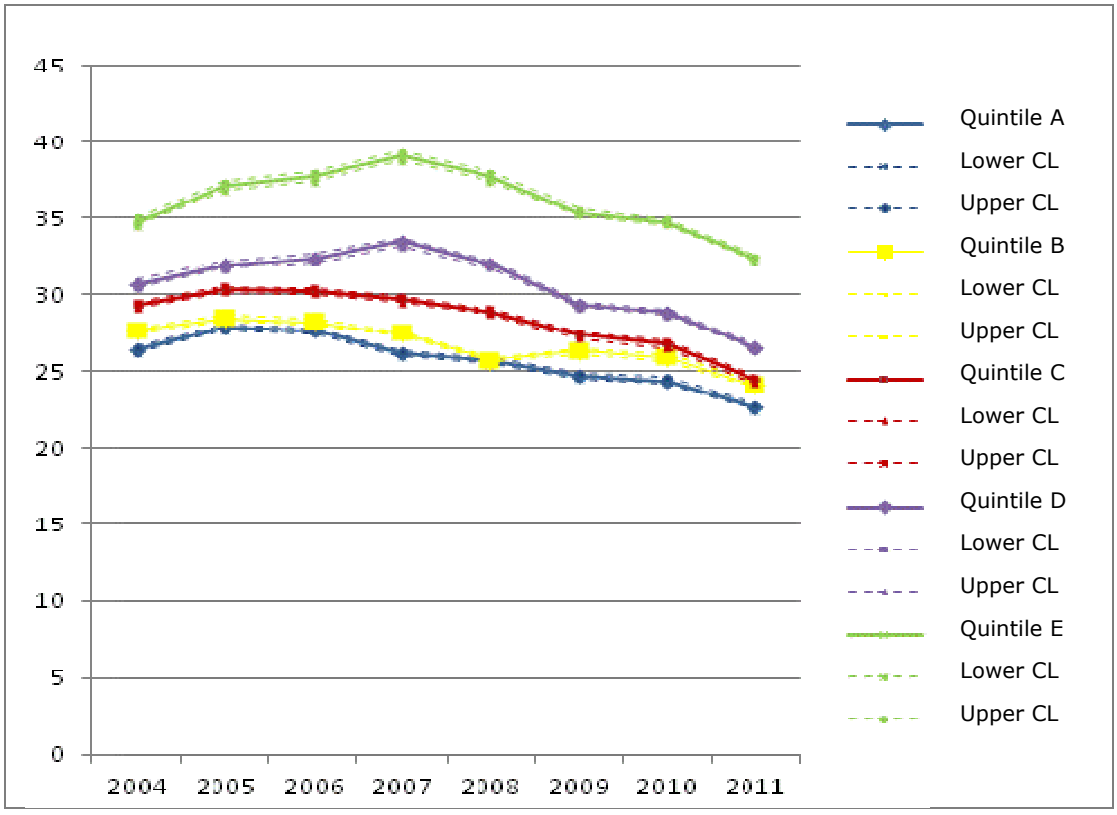


Figure 106: History Raw Score by SES Quintile

Appendix M. Weighted Correlations of NEM and PSU Subtest Raw Score

Table 213: Weighted Correlations of NEM and PSU Subtest Raw Score

Year	PSU Subtest	Weighted r
2004	Language and Communication and Mathematics	0.46
2005	Language and Communication and Mathematics	0.56
2006	Language and Communication and Mathematics	0.59
2007	Language and Communication and Mathematics	0.59
2008	Language and Communication and Mathematics	0.60
2009	Language and Communication and Mathematics	0.59
2010	Language and Communication and Mathematics	0.58
2011	Language and Communication and Mathematics	0.55
2004	Language and Communication	0.34
2005	Language and Communication	0.44
2006	Language and Communication	0.48
2007	Language and Communication	0.47
2008	Language and Communication	0.47
2009	Language and Communication	0.46
2010	Language and Communication	0.46
2011	Language and Communication	0.41
2004	Mathematics	0.43
2005	Mathematics	0.50
2006	Mathematics	0.53
2007	Mathematics	0.53
2008	Mathematics	0.54
2009	Mathematics	0.55
2010	Mathematics	0.52
2011	Mathematics	0.50
2004	History and Social Science	0.25
2005	History and Social Science	0.31
2006	History and Social Science	0.32
2007	History and Social Science	0.32
2008	History and Social Science	0.34
2009	History and Social Science	0.33
2010	History and Social Science	0.32
2011	History and Social Science	0.31
2004	Science	0.48
2005	Science	0.55
2006	Science	0.56
2007	Science	0.55
2008	Science	0.57
2009	Science	0.55
2010	Science	0.55
2011	Science	0.57

Appendix N. Results of Hierarchical Linear Modeling Analysis for Raw Scores

Table 214: Results of Hierarchical Linear Modeling Analysis – Language and Communication and Mathematics Raw Score

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		72.06	0.57	2414	127.24	0.00
NEM		0.05	0.00	231144	32.66	0.00
School SES		1.02	0.03	231144	35.96	0.00
% Female		-2.72	0.91	231144	-3.01	0.00
Region	Central	3.55	0.29	231144	12.34	0.00
	North	-7.35	0.35	231144	-21.20	0.00
	South	0.00				
School Type	Private	21.84	0.39	231144	55.59	0.00
	Subsidized	6.46	0.29	231144	21.99	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	7.43	0.24	231144	31.45	0.00
	Technical-Professional	0.00				
SES*NEM		0.00	0.00	231144	8.05	0.00
NEM*%*Female		0.02	0.00	231144	11.34	0.00
NEM*Region	Central	-0.01	0.00	231144	-15.11	0.00
	North	-0.01	0.00	231144	-6.38	0.00
	South	0.00				
NEM*Type	Private	0.04	0.00	231144	25.92	0.00
	Subsidized	0.02	0.00	231144	16.93	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.07	0.00	231144	55.89	0.00
	Technical-Professional	0.00				

Table 215: Results of Hierarchical Linear Modeling Analysis – Language and Communication Raw Score

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		41.86	0.24	2414	175.07	0.00
NEM		0.02	0.00	231143	20.02	0.00
School SES		0.52	0.02	231143	32.51	0.00
% Female		0.77	0.36	231143	2.12	0.03
Region	Central	2.67	0.14	231143	19.28	0.00
	North	-2.60	0.18	231143	-14.79	0.00
	South	0.00				
School Type	Private	8.05	0.19	231143	42.27	0.00
	Subsidized	2.65	0.14	231143	18.41	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	2.57	0.12	231143	20.60	0.00
	Technical-Professional	0.00				
SES*NEM		0.00	0.00	231143	6.06	0.00
NEM*%*Female			0.00	231143	9.43	0.00
NEM*Region	Central	-0.01	0.00	231143	-16.16	0.00
	North	0.00	0.00	231143	-3.25	0.00
	South	0.00				
NEM*Type	Private	0.02	0.00	231143	27.60	0.00
	Subsidized	0.01	0.00	231143	17.98	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.03	0.00	231143	38.13	0.00
	Technical-Professional	0.00				

Table 216: Results of Hierarchical Linear Modeling Analysis – Mathematics Raw Score

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		30.14	0.37	2414	81.77	0.00
NEM		0.03	0.00	231140	30.99	0.00
School SES		0.50	0.02	231140	25.60	0.00
% Female		-3.53	0.58	231140	-6.05	0.00
Region	Central	0.90	0.19	231140	4.65	0.00
	North	-4.78	0.23	231140	-20.37	0.00
	South	0.00				
School Type	Private	13.99	0.26	231140	53.03	0.00
	Subsidized	3.74	0.20	231140	18.88	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	5.02	0.16	231140	31.15	0.00
	Technical-Professional	0.00				
SES*NEM		0.00	0.00	231140	6.79	0.00
NEM*%*Female		0.01	0.00	231140	8.78	0.00
NEM*Region	Central	-0.01	0.00	231140	-8.67	0.00
	North	-0.01	0.00	231140	-6.60	0.00
	South	0.00				
NEM*Type	Private	0.01	0.00	231140	15.08	0.00
	Subsidized	0.01	0.00	231140	9.87	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.04	0.00	231140	50.08	0.00
	Technical-Professional	0.00				

Table 217: Results of Hierarchical Linear Modeling Analysis – Science Raw Score

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		25.70	0.39	2051	66.51	0.00
NEM		0.03	0.00	157513	22.28	0.00
School SES		0.46	0.02	157513	19.11	0.00
% Female		-1.27	0.60	157513	-2.13	0.03
Region	Central	2.99	0.23	157513	13.27	0.00
	North	-1.85	0.28	157513	-6.68	0.00
	South	0.00				
School Type	Private	13.48	0.31	157513	42.99	0.00
	Subsidized	3.04	0.23	157513	12.94	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	4.75	0.21	157513	22.17	0.00
	Technical-Professional	0.00				
SES*NEM		0.00	0.00	157513	7.42	0.00
NEM*%*Female		0.00	0.00	157513	1.37	0.17
NEM*Region	Central	-0.01	0.00	157513	-7.57	0.00
	North	-0.01	0.00	157513	-11.89	0.00
	South	0.00				
NEM*Type	Private	0.03	0.00	157513	25.95	0.00
	Subsidized	0.01	0.00	157513	9.58	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.05	0.00	157513	46.96	0.00
	Technical-Professional	0.00				

Table 218: Results of Hierarchical Linear Modeling Analysis – History and Social Sciences Raw Score

Effect	Region	Estimate	S.E.	DF	t-Value	Prob
Y_{00}		39.19	0.32	2050	121.10	0.00
NEM		0.00	0.00	137701	2.88	0.00
School SES		0.31	0.03	137701	11.61	0.00
% Female		-2.71	0.46	137701	-5.87	0.00
Region	Central	3.58	0.20	137701	18.21	0.00
	North	-2.78	0.26	137701	-10.81	0.00
	South	0.00				
School Type	Private	8.21	0.27	137701	30.08	0.00
	Subsidized	2.24	0.20	137701	10.93	0.00
	Municipal	0.00				
Curricular Branch	Scientific-Humanistic	1.97	0.18	137701	10.66	0.00
	Technical-Professional	0.00				
SES*NEM		0.00	0.00	137701	4.71	0.00
NEM*%*Female			0.00	137701	8.97	0.00
NEM*Region	Central	0.00	0.00	137701	-4.26	0.00
	North	0.00	0.00	137701	-0.20	0.84
	South	0.00				
NEM*Type	Private	0.03	0.00	137701	25.09	0.00
	Subsidized	0.01	0.00	137701	11.74	0.00
	Municipal	0.00				
NEM*Branch	Scientific-Humanistic	0.02	0.00	137701	18.18	0.00
	Technical-Professional	0.00				

Appendix O. Descriptive Statistics for Unrestricted Predictors

Table 219: Descriptive Statistics for Unrestricted Predictors

Year	N	Mean	S.D.	Variance
NEM				
2004	151198	559.75	101.04	10209.62
2005	165764	562.49	99.96	9992.33
2006	172927	559.40	99.68	9936.52
2007	206120	549.44	100.52	10103.90
2008	213184	546.17	100.15	10029.05
2009	238744	540.32	100.30	10060.89
2010	248545	536.11	99.90	9979.94
2011	248807	535.81	99.88	9976.68
2012	228722	539.35	99.74	9948.79
Language and Communication				
2004	153383	501.03	121.76	14824.35
2005	169376	500.27	109.17	11919.09
2006	176314	500.23	109.04	11890.29
2007	211261	500.59	109.02	11884.94
2008	216892	500.50	109.15	11914.36
2009	242130	500.61	108.99	11879.35
2010	251634	500.64	108.92	11863.24
2011	250758	501.04	108.34	11736.52
2012	231140	500.69	108.99	11878.32
History and Social Science				
2004	106443	500.38	107.51	11558.59
2005	112857	500.23	109.11	11905.73
2006	116097	500.17	109.10	11903.52
2007	137717	500.25	109.26	11937.84
2008	138974	500.13	109.29	11944.97
2009	152217	500.30	109.40	11968.22
2010	157248	500.32	109.49	11987.78
2011	154790	500.41	109.55	12000.34
2012	140114	500.94	110.93	12305.19
Mathematics				
2004	153383	499.99	109.46	11981.25
2005	169376	500.55	110.14	12129.81
2006	176314	500.61	110.24	12153.68
2007	211261	500.31	109.55	12000.53
2008	216892	500.36	109.87	12072.09
2009	242130	500.20	109.59	12009.52
2010	251634	500.79	110.77	12270.17
2011	250758	501.07	111.27	12380.78
2012	231140	500.36	109.73	12040.97
Science				
2004	84136	500.58	99.78	9955.74
2005	93935	500.22	109.27	11939.90
2006	99342	500.30	109.16	11915.97

2007	113024	500.55	109.11	11904.74
2008	120219	500.32	109.46	11981.25
2009	137996	500.47	109.22	11928.94
2010	141325	500.60	109.08	11897.63
2011	139783	500.52	109.47	11984.60
2012	132969	500.56	109.28	11943.20
Rank*				
2003	164381	53.69	29.08	845.45
2004	187169	52.12	29.01	841.87
2005	200204	51.46	28.97	839.35
2006	202522	51.32	28.96	838.75
2007	206183	51.27	28.90	835.50
2008	207307	51.29	28.91	835.62
2009	203656	51.25	28.90	835.18
2010	203123	51.28	28.90	835.17

(Note: Information for Rank variable available from MINEDUC data file spanned 2003-2010 admission process)

Appendix P. Prediction Validity by the Type of Career - First Year Grade Point Average (FYGPA)

Table 220: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Administración)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	19	24	1214	1114	656	0.14	0.21	0.10	0.18	0.32	0.32
2005	18	27	1780	1603	896	0.14	0.19	0.09	0.10	0.32	0.31
2006	20	30	2151	1872	1117	0.20	0.19	0.13	0.18	0.30	0.30
2007	20	30	2556	2169	1312	0.17	0.18	0.07	0.16	0.30	0.26
2008	20	29	2922	2420	1562	0.11	0.13	0.06	0.17	0.33	0.30
2009	21	30	3054	2396	1744	0.15	0.28	0.07	0.21	0.34	0.28
2010	19	27	2942	2243	1659	0.19	0.18	0.10	0.21	0.34	0.26
2011	1*	1	278	207	166	0.14	-0.07	0.04	0.35	0.42	N/A

Table 221: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Administración_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	12	382	354	180	0.10	0.11	0.06	0.03	0.32	0.42
2005	12	15	572	526	218	0.22	0.23	0.08	0.18	0.37	0.43
2006	15	19	903	790	406	0.19	0.16	0.11	0.11	0.28	0.28
2007	16	20	1132	979	485	0.10	0.17	0.07	0.11	0.34	0.32
2008	16	19	1016	862	493	0.19	0.21	0.16	0.16	0.24	0.25
2009	16	19	1080	903	555	0.17	0.18	0.06	0.17	0.29	0.31
2010	15	17	1004	809	530	0.06	0.28	-0.02	0.17	0.31	0.26
2011	1*	1	117	81	75	-0.23	0.12	-0.41	0.52	0.18	N/A

Table 222: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Administración_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	4	98	94	28	0.15	0.39	0.26	0.51	0.13	0.06
2005	4	5	168	154	42	0.17	0.31	0.15	0.23	0.24	0.26
2006	4	8	298	260	107	0.29	0.27	0.17	0.45	0.40	0.37
2007	5	9	320	278	97	0.33	0.37	0.21	0.18	0.36	0.24
2008	6	11	420	345	157	0.33	0.27	0.27	0.14	0.40	0.26
2009	6	10	396	307	177	0.21	0.45	0.10	0.15	0.46	0.32
2010	6	10	428	330	183	0.36	0.35	0.21	0.39	0.49	0.38
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

*Note: Year 2011 is shown in compliance with contract expectations. Caution is recommended when making inferences given the small number of universities.

Table 223: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Agro)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	8	11	521	198	521	0.07	0.22	0.01	0.20	0.33	0.28
2005	8	13	797	258	792	0.22	0.22	0.08	0.27	0.29	0.27
2006	8	13	925	317	920	0.06	0.19	0.17	0.23	0.31	0.32
2007	8	14	994	354	984	0.14	0.21	0.08	0.27	0.25	0.21
2008	8	14	975	350	971	0.06	0.16	0.13	0.18	0.27	0.24
2009	8	14	949	321	936	0.17	0.34	0.07	0.30	0.29	0.22
2010	8	13	917	338	894	0.18	0.27	0.14	0.27	0.22	0.15
2011	1*	1	19	-	19	-0.17	-0.36	-	-0.56	0.10	N/A

Table 224: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Agro_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	51	30	44	-0.06	0.10	0.08	0.22	0.02	0.41
2005	3	4	101	66	77	0.03	0.00	0.04	0.17	0.29	0.19
2006	3	3	85	54	63	0.20	-0.06	0.01	0.39	0.29	0.20
2007	3	4	110	59	91	0.31	0.29	0.04	0.31	0.26	0.23
2008	3	4	95	55	80	0.10	0.36	0.07	0.38	0.10	0.08
2009	3	4	116	58	102	0.36	0.38	0.05	0.39	0.16	0.01
2010	3	4	94	42	84	-0.04	0.19	-0.20	0.07	0.33	0.32
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 225: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Arquitectura)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	11	596	577	263	0.03	0.18	0.14	0.08	0.32	0.30
2005	12	12	804	756	345	0.01	0.25	0.11	0.19	0.30	0.26
2006	12	12	903	821	420	0.07	0.16	0.10	0.08	0.23	0.22
2007	12	12	904	835	388	0.05	0.25	0.11	0.07	0.32	0.27
2008	14	15	963	853	459	0.09	0.20	0.03	0.09	0.32	0.25
2009	14	15	999	854	519	0.10	0.14	0.10	0.16	0.27	0.22
2010	12	13	731	584	388	0.05	0.17	0.00	0.19	0.24	0.16
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

*Note: Year 2011 is shown in compliance with contract expectations. Caution is recommended when making inferences given the small number of universities.

Table 226: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Arte_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	5	190	187	36	0.15	0.14	0.11	0.43	0.25	0.24
2005	6	8	311	311	40	0.10	0.20	0.05	0.88	0.19	0.24
2006	9	10	392	387	78	-0.04	0.13	0.03	-0.13	0.26	0.27
2007	9	11	410	404	72	0.17	0.03	0.15	0.48	0.23	0.21
2008	8	9	373	371	62	0.06	0.10	0.08	-0.14	0.32	0.21
2009	9	11	431	426	60	0.00	0.12	0.14	0.38	0.36	0.27
2010	9	12	394	377	77	0.24	0.29	0.10	0.38	0.31	0.28
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 227: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Arte_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	4	98	98	11	0.31	0.28	0.22	0.20	0.43	0.38
2005	5	9	235	224	59	0.22	0.30	0.00	0.41	0.31	0.30
2006	5	7	224	223	34	0.21	0.17	0.13	0.38	0.13	0.12
2007	5	9	243	234	50	0.16	0.23	0.01	0.41	0.33	0.20
2008	5	9	262	251	63	0.18	0.18	0.26	0.10	0.22	0.15
2009	6	10	292	273	71	0.12	0.35	0.16	0.17	0.23	0.24
2010	5	7	156	156	18	0.26	0.22	0.29	0.47	0.27	0.23
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 228: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ciencias)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	7	10	399	135	376	0.06	0.17	-0.03	0.30	0.29	0.25
2005	7	12	505	114	492	0.13	0.24	0.11	0.24	0.30	0.22
2006	8	14	604	156	596	0.18	0.26	0.02	0.20	0.18	0.25
2007	8	15	710	161	700	0.14	0.21	0.31	0.24	0.23	0.23
2008	8	17	855	223	803	0.27	0.22	0.43	0.22	0.21	0.16
2009	7	17	1017	313	970	0.22	0.28	0.30	0.37	0.22	0.21
2010	7	17	962	255	926	0.29	0.28	0.14	0.39	0.31	0.20
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 229: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ciencias_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	21	630	150	630	0.13	0.31	-0.09	0.21	0.25	0.26
2005	10	22	870	199	868	0.13	0.14	0.22	0.25	0.25	0.20
2006	11	23	993	219	993	0.19	0.21	0.20	0.22	0.18	0.17
2007	12	24	1064	202	1063	0.18	0.19	0.06	0.21	0.23	0.21
2008	12	24	1098	225	1097	0.14	0.30	0.11	0.31	0.26	0.17
2009	12	25	1077	233	1070	0.12	0.38	0.22	0.24	0.23	0.17
2010	11	22	986	210	967	0.15	0.28	0.07	0.29	0.20	0.17
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 230: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ciencias_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	6	171	70	166	0.07	0.27	-0.10	0.09	0.32	0.23
2005	5	8	283	108	269	0.29	0.31	0.40	0.31	0.26	0.21
2006	6	10	322	127	299	0.21	0.13	-0.18	0.21	0.22	0.27
2007	6	10	350	124	332	0.25	0.40	-0.13	0.28	0.24	0.21
2008	5	9	279	109	256	0.33	0.36	0.11	0.38	0.25	0.21
2009	5	8	246	70	239	0.20	0.25	0.54	0.26	0.23	0.19
2010	5	9	243	63	232	0.21	0.45	-0.06	0.39	0.39	0.34
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 231: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ciencias_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	17	-	17	0.35	0.52	-	0.66	0.24	0.09
2005	2	2	54	24	54	0.38	0.50	0.43	0.51	0.27	0.27
2006	2	2	46	-	46	0.19	0.20	-	-0.17	0.22	0.21
2007	2	4	104	-	104	0.16	0.31	-	0.25	-0.03	0.09
2008	2	4	143	51	143	0.16	0.40	0.28	0.18	0.13	0.17
2009	2	4	153	51	153	0.13	0.45	0.12	0.38	0.31	0.17
2010	2	4	158	49	158	0.03	0.39	0.00	0.31	0.09	0.11
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 232: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ciencias_Sociales_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	13	21	745	707	283	0.10	0.21	0.05	0.15	0.35	0.33
2005	15	26	1199	1121	382	0.08	0.06	0.11	-0.09	0.25	0.17
2006	18	29	1489	1349	547	0.13	0.20	0.10	0.20	0.35	0.29
2007	17	28	1525	1399	542	0.10	0.20	0.03	0.09	0.28	0.24
2008	18	29	1561	1417	556	0.07	0.18	0.07	0.05	0.33	0.25
2009	18	30	1594	1433	568	0.19	0.20	0.17	0.15	0.31	0.27
2010	17	29	1513	1320	594	0.17	0.25	0.10	0.27	0.35	0.27
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 233: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ciencias_Sociales_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	6	13	491	491	112	0.11	0.16	0.17	-0.08	0.19	0.24
2005	10	18	806	805	128	0.04	0.17	-0.01	0.24	0.33	0.36
2006	12	20	1005	1004	172	0.07	0.10	0.00	0.04	0.29	0.28
2007	13	28	1332	1325	212	0.10	0.13	0.08	0.25	0.30	0.26
2008	13	29	1401	1396	220	0.12	0.24	0.07	0.24	0.28	0.23
2009	13	29	1574	1568	319	0.10	0.13	0.07	0.15	0.26	0.26
2010	12	27	1368	1361	264	0.09	0.24	0.09	-0.08	0.30	0.24
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 234: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ciencias_Sociales_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	11	262	261	70	0.18	0.18	0.06	0.06	0.29	0.30
2005	11	12	434	431	102	-0.01	-0.02	0.10	-0.01	0.33	0.27
2006	12	13	572	570	128	0.07	0.14	-0.07	0.05	0.27	0.21
2007	12	14	617	614	140	0.10	0.23	0.07	-0.19	0.27	0.24
2008	11	13	573	571	135	0.13	0.06	-0.02	0.15	0.32	0.22
2009	12	14	638	636	184	0.10	0.22	0.08	0.12	0.28	0.21
2010	11	12	558	553	125	0.05	-0.04	0.06	-0.21	0.19	0.19
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 235: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Comunicaciones)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	23	22	9	-0.02	-0.25	-0.03	-0.24	0.21	0.47
2005	1	1	36	34	-	-0.24	0.02	-0.09	-	0.23	0.35
2006	2	2	73	68	28	-0.19	0.07	-0.14	0.24	0.25	0.24
2007	2	2	79	75	-	-0.10	0.14	0.15	-	0.28	0.41
2008	2	2	71	68	-	-0.12	0.03	0.00	-	0.08	0.23
2009	2	2	54	49	-	-0.03	-0.13	-0.48	-	0.32	0.26
2010	2	2	72	70	-	-0.11	0.19	-0.03	-	-0.01	-0.16
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 236: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Construcción)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	11	396	166	391	0.05	0.20	0.23	0.18	0.24	0.24
2005	13	13	546	222	536	0.20	0.18	0.08	0.24	0.22	0.18
2006	13	13	668	282	649	0.02	0.22	0.03	0.12	0.11	0.13
2007	15	15	798	314	773	0.12	0.28	0.09	0.15	0.22	0.25
2008	14	14	808	325	778	0.06	0.26	-0.02	0.22	0.15	0.12
2009	14	14	847	323	814	0.08	0.33	0.01	0.26	0.21	0.13
2010	11	11	690	279	659	0.11	0.25	0.13	0.19	0.14	0.10
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 237: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Derecho)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	15	18	1195	1191	325	0.08	0.18	0.10	0.14	0.32	0.29
2005	15	19	1571	1568	339	0.11	0.16	0.12	0.13	0.34	0.32
2006	15	19	1801	1789	380	0.13	0.18	0.12	0.18	0.33	0.29
2007	15	19	1983	1973	378	0.18	0.24	0.15	0.17	0.32	0.27
2008	15	19	2074	2070	363	0.06	0.18	0.11	0.16	0.35	0.33
2009	14	19	2054	2032	401	0.07	0.24	0.14	0.09	0.34	0.32
2010	14	19	2014	1965	432	0.15	0.21	0.21	0.20	0.32	0.26
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 238: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Diseño)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	9	11	444	435	93	0.07	0.18	0.09	0.17	0.28	0.23
2005	11	15	699	669	166	-0.05	0.21	0.02	-0.13	0.31	0.26
2006	12	18	944	886	263	-0.01	0.16	0.00	0.11	0.30	0.27
2007	11	17	968	892	280	0.05	0.06	0.02	-0.02	0.21	0.18
2008	12	18	981	892	278	0.09	0.21	0.06	0.02	0.34	0.26
2009	12	18	945	872	272	0.05	0.17	0.01	0.28	0.28	0.22
2010	11	15	620	567	163	0.10	0.25	0.14	0.17	0.22	0.17
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 239: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Educación)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	14	28	576	576	159	0.00	0.10	0.02	0.14	0.30	0.27
2005	15	50	1393	1386	322	0.01	0.14	-0.03	0.15	0.27	0.24
2006	16	55	1757	1751	363	0.11	0.15	0.00	0.10	0.24	0.20
2007	16	60	2020	1996	394	0.07	0.18	0.00	0.07	0.25	0.18
2008	16	58	1973	1956	402	0.07	0.15	0.05	-0.07	0.25	0.23
2009	17	58	1968	1930	459	0.09	0.21	0.05	0.01	0.27	0.20
2010	16	53	1798	1736	423	0.05	0.11	0.02	0.30	0.29	0.24
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 240: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Educación_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	12	28	665	627	255	0.04	0.19	0.02	0.06	0.30	0.25
2005	13	44	1449	1346	583	0.00	0.12	0.04	0.16	0.28	0.26
2006	14	50	1850	1676	773	0.11	0.12	0.08	0.05	0.25	0.23
2007	16	52	2062	1851	840	0.03	0.05	0.04	0.03	0.29	0.27
2008	16	50	2065	1846	888	0.04	0.10	0.07	0.17	0.26	0.20
2009	15	50	2155	1963	911	0.08	0.23	0.05	0.15	0.25	0.19
2010	15	50	1921	1724	831	0.17	0.20	0.11	0.10	0.24	0.15
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 241: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Educación_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	14	303	143	301	0.15	0.24	0.06	0.20	0.26	0.28
2005	11	26	806	327	804	0.12	0.26	0.13	0.28	0.23	0.23
2006	13	33	1064	440	1053	0.16	0.28	0.17	0.18	0.24	0.20
2007	12	32	1101	431	1091	0.11	0.05	0.14	0.04	0.24	0.21
2008	12	28	1006	354	998	0.16	0.18	-0.05	0.23	0.27	0.22
2009	12	32	1114	416	1108	0.21	0.23	0.24	0.18	0.27	0.23
2010	11	28	953	320	944	0.16	0.28	0.02	0.29	0.26	0.23
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 242: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Educación_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	8	10	249	139	201	0.09	0.15	0.01	0.13	0.27	0.21
2005	8	13	483	302	379	0.08	0.14	0.07	0.09	0.25	0.16
2006	11	17	676	433	536	0.14	0.13	0.05	0.03	0.36	0.28
2007	11	17	697	393	546	0.11	0.13	0.03	0.10	0.32	0.26
2008	11	17	713	371	575	0.13	0.16	0.09	0.19	0.28	0.26
2009	11	17	756	399	616	0.07	0.17	0.06	0.16	0.32	0.29
2010	10	16	719	359	580	0.10	0.20	-0.01	0.11	0.29	0.24
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 243: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (General)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	185	74	157	0.11	0.06	-0.12	0.21	0.15	0.20
2005	1	1	235	89	215	0.20	-0.27	0.13	0.08	0.18	0.09
2006	1	1	243	89	228	0.05	-0.04	-0.23	0.08	0.20	0.22
2007	1	1	263	82	244	0.06	0.03	0.01	0.07	0.06	-0.01
2008	1	1	273	85	257	0.08	-0.01	0.04	0.17	0.19	0.04
2009	1	1	263	91	238	-0.07	0.19	0.02	-0.02	0.14	0.15
2010	1	1	251	106	213	0.05	0.12	0.18	0.25	0.11	0.17
2011	1*	1	16	-	-	-0.71	-0.20	-	-	0.12	N/A

*Note: Year 2011 is shown in compliance with contract expectations. Caution is recommended when making inferences given the small number of universities.

Table 244: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Humanidades)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	5	8	267	267	53	0.08	0.08	0.18	0.02	0.43	0.23
2005	8	12	487	483	-	0.06	0.21	0.11	-	0.18	0.19
2006	9	14	536	532	104	0.12	0.11	0.07	0.30	0.30	0.33
2007	9	14	578	570	87	0.06	0.28	0.18	0.68	0.29	0.22
2008	10	15	535	529	74	0.23	0.21	0.21	0.87	0.32	0.23
2009	10	15	530	522	-	0.03	0.19	0.07	-	0.29	0.22
2010	8	12	385	384	-	0.18	0.29	0.13	-	0.38	0.30
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 245: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ingeniería_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	21	80	3729	1173	3713	0.15	0.26	0.05	0.25	0.27	0.27
2005	21	93	5168	1727	5094	0.18	0.26	0.10	0.28	0.26	0.24
2006	22	102	5907	2085	5788	0.17	0.27	0.08	0.26	0.24	0.19
2007	22	112	6675	2304	6498	0.16	0.29	0.11	0.24	0.25	0.22
2008	22	115	6921	2303	6740	0.19	0.26	0.12	0.31	0.25	0.22
2009	22	118	7501	2492	7300	0.17	0.31	0.06	0.29	0.25	0.22
2010	20	97	6358	2038	6179	0.18	0.31	0.13	0.31	0.27	0.24
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 246: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ingeniería_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	30	728	233	718	0.06	0.24	-0.06	0.12	0.16	0.16
2005	13	45	1374	586	1276	0.12	0.26	0.06	0.20	0.21	0.18
2006	15	58	1793	714	1660	0.08	0.23	0.04	0.17	0.24	0.22
2007	15	57	1951	738	1779	0.05	0.23	0.06	0.15	0.20	0.18
2008	17	58	1957	738	1761	0.18	0.22	0.04	0.18	0.24	0.19
2009	16	63	2090	808	1908	0.15	0.31	0.12	0.21	0.26	0.23
2010	13	57	2019	780	1822	0.07	0.28	0.05	0.25	0.23	0.16
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 247: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Ingeniería_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	5	120	33	120	0.09	0.31	-0.03	0.22	0.27	0.38
2005	3	5	161	-	161	0.29	0.17	-	0.30	0.33	0.23
2006	4	8	301	78	300	0.24	0.14	0.53	0.25	0.33	0.33
2007	6	12	470	129	469	0.23	0.26	0.14	0.35	0.32	0.31
2008	6	12	458	107	456	0.25	0.33	0.06	0.39	0.33	0.28
2009	6	12	454	114	453	0.21	0.27	0.05	0.37	0.27	0.20
2010	6	12	432	119	432	0.12	0.35	0.07	0.32	0.31	0.23
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 248: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Mar)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	5	6	124	64	124	0.18	0.46	0.13	0.27	0.22	0.19
2005	8	11	282	103	275	0.25	0.31	0.04	0.27	0.34	0.20
2006	8	10	283	120	274	0.09	0.33	0.01	0.26	0.25	0.18
2007	7	10	275	93	269	0.21	0.39	-0.07	0.32	0.25	0.22
2008	8	12	360	116	348	0.27	0.32	-0.35	0.26	0.27	0.16
2009	8	12	274	106	264	0.13	0.30	0.02	0.18	0.36	0.30
2010	8	10	242	86	225	0.28	0.40	0.31	0.39	0.32	0.21
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 249: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Periodismo)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	12	12	457	455	89	0.13	0.13	0.14	0.35	0.22	0.28
2005	12	12	597	597	94	0.09	0.20	0.15	-0.28	0.34	0.35
2006	15	16	763	759	116	0.14	0.07	0.12	-0.25	0.37	0.26
2007	14	14	762	761	114	0.17	0.27	0.09	-0.05	0.23	0.24
2008	14	15	668	665	116	0.04	0.17	0.20	0.47	0.35	0.28
2009	13	14	671	665	113	0.02	0.22	0.00	-0.36	0.30	0.27
2010	13	13	581	554	127	0.02	0.21	0.00	0.52	0.34	0.33
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 250: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Salud_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	12	20	895	199	894	0.02	0.12	0.03	0.25	0.29	0.22
2005	14	26	1477	327	1476	0.07	0.03	0.01	0.18	0.28	0.31
2006	14	26	1785	412	1782	0.09	0.18	0.12	0.28	0.23	0.24
2007	14	26	1841	425	1837	0.05	0.15	0.05	0.19	0.24	0.22
2008	15	28	1861	390	1858	-0.05	0.13	0.20	0.28	0.22	0.20
2009	15	30	2111	478	2108	-0.01	0.15	-0.04	0.24	0.33	0.30
2010	14	28	1891	434	1867	0.08	0.10	0.26	0.31	0.30	0.27
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 251: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Salud_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	20	43	1341	376	1337	0.07	0.21	0.12	0.23	0.26	0.28
2005	19	45	2159	589	2151	0.17	0.19	0.19	0.26	0.24	0.25
2006	20	47	2590	689	2574	0.16	0.14	0.21	0.23	0.27	0.23
2007	19	48	2707	660	2686	0.14	0.16	0.15	0.20	0.30	0.22
2008	19	48	2791	701	2773	0.09	0.15	0.06	0.22	0.20	0.19
2009	19	51	2996	712	2977	0.14	0.18	0.03	0.24	0.25	0.23
2010	19	53	3083	841	2983	0.15	0.16	0.05	0.25	0.28	0.27
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 252: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Salud_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	17	536	148	530	0.07	0.18	0.18	0.19	0.25	0.28
2005	13	22	1027	274	1022	0.11	0.20	0.18	0.19	0.30	0.26
2006	14	23	1220	317	1213	0.16	0.26	0.05	0.22	0.31	0.26
2007	14	23	1303	303	1295	0.14	0.22	0.01	0.25	0.30	0.20
2008	14	23	1320	283	1315	0.11	0.21	0.00	0.27	0.20	0.19
2009	14	23	1282	294	1269	0.14	0.22	0.05	0.20	0.25	0.22
2010	14	25	1227	288	1199	0.18	0.25	0.11	0.28	0.22	0.18
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 253: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Técnico_Administración)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	3	112	105	53	0.04	0.08	0.00	0.07	0.25	0.33
2005	5	6	239	214	80	-0.03	0.15	0.02	0.32	0.26	0.22
2006	5	6	297	247	112	0.11	0.17	0.01	-0.02	0.24	0.34
2007	5	7	289	252	99	0.19	0.25	0.03	0.41	0.31	0.25
2008	6	8	325	275	121	0.07	0.20	0.18	0.21	0.23	0.32
2009	6	7	282	231	111	0.28	0.19	0.10	0.19	0.08	0.12
2010	7	9	354	287	131	0.15	0.21	-0.02	0.12	0.35	0.32
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 254: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Técnico_Agro)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	15	10	15	0.07	0.34	0.08	0.11	0.36	0.23
2005	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2006	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2007	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2008	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2009	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 255: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Técnico_Ciencias)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	27	-	27	0.00	-0.15	-	0.11	0.42	0.25
2005	2	3	76	30	74	0.04	0.38	0.06	0.13	0.39	0.36
2006	2	3	90	-	89	-0.07	0.27	-	-0.04	0.03	0.00
2007	2	3	77	35	76	0.13	0.20	0.46	0.22	0.40	0.26
2008	3	4	123	42	121	0.22	-0.04	0.09	0.07	0.25	0.22
2009	3	4	114	-	111	0.16	0.31	-	0.24	0.13	0.28
2010	2	2	49	-	49	-0.17	-0.12	-	0.10	0.14	0.33
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 256: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Técnico_Diseño)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	16	16	-	-0.05	0.13	0.33	-	0.30	0.52
2005	2	2	45	41	-	-0.04	0.21	0.01	-	0.29	0.09
2006	3	3	71	55	42	0.08	0.23	-0.21	-0.09	0.15	0.25
2007	3	3	69	52	35	0.01	0.07	-0.13	-0.01	0.41	0.24
2008	3	3	68	49	34	0.25	0.25	-0.02	0.25	0.33	0.23
2009	3	3	69	40	49	0.43	0.29	0.08	0.59	0.03	-0.01
2010	1	1	28	18	16	-0.16	0.30	0.13	0.16	-0.05	-0.29
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 257: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Técnico_Educación)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	27	18	22	-0.19	0.36	-0.15	0.29	0.08	-
2005	1	1	22	13	19	-0.11	0.07	-0.53	-0.19	0.11	-0.21
2006	1	1	24	16	19	0.31	0.29	0.25	0.57	0.16	0.12
2007	1	1	18	-	-	-0.15	-0.17	-	-	0.26	0.07
2008	1	1	20	-	15	-0.24	0.50	-	-0.28	0.40	0.37
2009	1	1	27	19	21	0.30	-0.33	-0.28	0.02	0.32	0.35
2010	1	1	25	19	18	0.22	0.15	-0.07	0.40	-0.04	0.05
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 258: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Técnico_Idioma)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	60	60	14	-0.01	0.12	0.11	0.06	0.33	0.49
2005	3	4	134	129	49	0.23	0.17	0.16	0.03	0.27	0.14
2006	3	4	150	145	49	0.11	-0.06	0.12	0.06	0.21	0.17
2007	3	4	156	147	52	0.04	0.29	-0.08	0.29	0.23	0.12
2008	3	3	139	135	43	0.06	0.38	0.07	0.28	0.30	0.11
2009	3	3	128	117	40	0.20	0.38	0.13	0.10	0.17	0.19
2010	3	3	99	91	39	0.14	0.04	0.11	-0.13	0.18	0.13
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 259: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Técnico_Ingeniería)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	7	21	497	186	475	0.10	0.30	0.12	0.18	0.21	0.23
2005	7	41	1202	525	1086	0.18	0.32	0.10	0.24	0.15	0.15
2006	6	47	1442	657	1283	0.06	0.31	0.04	0.15	0.14	0.16
2007	7	48	1574	676	1378	0.14	0.26	0.05	0.21	0.11	0.07
2008	6	48	1843	819	1565	0.07	0.27	0.03	0.16	0.18	0.16
2009	6	49	1818	799	1586	0.04	0.26	0.03	0.16	0.13	0.12
2010	4	22	888	256	847	0.05	0.25	-0.05	0.11	0.03	0.02
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 260: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University FYGPA by Admission Year (Veterinaria)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	5	264	73	264	0.19	0.31	0.11	0.32	0.32	0.28
2005	4	5	404	97	401	0.19	0.28	0.18	0.42	0.22	0.16
2006	4	5	412	98	410	0.13	0.18	-0.08	0.30	0.25	0.29
2007	4	5	459	113	455	0.13	0.24	-0.10	0.12	0.25	0.20
2008	4	5	443	107	441	0.12	0.20	0.10	0.29	0.25	0.22
2009	4	5	421	93	416	0.11	0.19	0.04	0.28	0.28	0.23
2010	4	5	408	113	392	0.27	0.28	0.01	0.30	0.27	0.21
2011	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Appendix Q. Prediction Validity by the Type of Career – Second Year Grade Point Average (SYGPA)

Table 261: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Administración)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	20	23	1113	1020	619	0.14	0.09	0.13	0.06	0.35	0.37
2005	19	28	1696	1536	855	0.16	0.16	0.10	0.13	0.34	0.31
2006	20	29	1907	1673	1006	0.19	0.16	0.10	0.12	0.32	0.30
2007	20	30	2321	1984	1187	0.17	0.12	0.10	0.09	0.36	0.31
2008	20	28	2500	2100	1328	0.13	0.10	0.07	0.18	0.31	0.30
2009	20	29	2587	2047	1485	0.18	0.20	0.10	0.12	0.34	0.28
2010	2	3	320	246	191	0.11	-0.07	0.16	0.41	0.27	0.03

Table 262: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Administración_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	12	13	363	340	169	0.04	0.04	0.00	0.06	0.31	0.37
2005	14	16	584	541	231	0.21	0.17	0.13	0.11	0.32	0.39
2006	14	18	811	709	363	0.20	0.20	0.15	0.16	0.26	0.34
2007	15	19	1030	893	441	0.17	0.14	0.06	-0.03	0.34	0.37
2008	15	18	891	755	434	0.17	0.20	0.15	0.13	0.23	0.29
2009	16	19	990	834	511	0.14	0.15	0.13	0.12	0.33	0.34
2010	1	1	114	78	85	0.05	-0.07	-0.08	-0.11	0.22	0.18

Table 263: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Administración_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	4	88	84	26	-0.04	0.32	0.15	0.44	0.37	0.17
2005	4	5	141	129	36	0.06	0.07	0.07	0.35	0.19	0.25
2006	4	8	267	232	99	0.10	0.21	0.19	0.10	0.43	0.33
2007	5	9	272	233	88	0.27	0.26	0.20	0.15	0.40	0.27
2008	6	10	352	290	130	0.29	0.33	0.23	0.24	0.40	0.29
2009	6	10	340	262	150	0.20	0.41	0.09	0.19	0.44	0.34
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 264: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Agro)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	8	10	467	174	467	0.10	0.12	0.04	0.14	0.29	0.31
2005	8	12	718	227	715	0.18	0.16	0.14	0.21	0.31	0.25
2006	8	13	825	287	821	0.13	0.18	0.17	0.21	0.38	0.36
2007	8	14	860	296	850	0.23	0.36	0.13	0.29	0.29	0.26
2008	8	12	800	281	796	0.13	0.20	0.16	0.18	0.28	0.28
2009	8	14	814	273	804	0.16	0.26	0.06	0.23	0.35	0.31
2010	1	2	78	-	78	-0.17	0.16	-	0.06	0.33	0.24

Table 265: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Agro_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	46	27	40	-0.01	0.06	0.03	0.11	0.08	0.34
2005	3	4	92	57	71	0.05	0.20	-0.02	0.13	0.12	0.20
2006	3	3	76	50	56	0.14	0.03	-0.09	0.04	0.27	0.34
2007	3	3	82	41	66	0.11	0.25	0.39	0.48	0.00	0.12
2008	1	1	41	27	32	0.15	0.31	-0.06	0.29	-0.02	-0.14
2009	2	2	58	31	49	0.14	0.10	0.15	0.30	0.15	-0.02
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 266: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Arquitectura)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	10	520	509	229	0.07	0.19	0.07	0.18	0.28	0.27
2005	13	13	738	701	320	0.07	0.23	0.16	0.13	0.30	0.23
2006	12	12	793	724	367	0.00	0.13	0.06	0.09	0.22	0.19
2007	12	12	785	729	340	0.02	0.25	0.13	0.07	0.33	0.25
2008	12	13	760	669	372	0.14	0.16	0.07	0.07	0.32	0.22
2009	13	14	782	680	401	0.02	0.17	0.03	0.25	0.35	0.30
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 267: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Arte_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	5	176	175	34	0.09	0.09	0.12	0.56	0.27	0.24
2005	6	8	274	274	39	0.12	0.22	0.06	0.82	0.11	0.22
2006	9	10	350	347	69	-0.03	0.18	0.06	0.05	0.25	0.25
2007	7	9	326	323	52	0.21	0.05	0.19	0.04	0.34	0.27
2008	7	8	313	313	49	0.11	0.20	0.20	-0.17	0.32	0.23
2009	8	9	322	317	39	0.11	0.26	0.19	0.37	0.38	0.23
2010	1	1	16	16	-	0.06	0.05	0.24	-	0.34	0.53

Table 268: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Arte_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	3	77	77	8	0.32	0.21	0.28	-0.46	0.33	0.14
2005	5	9	213	203	51	0.26	0.30	0.01	0.47	0.40	0.31
2006	4	5	175	174	29	0.16	0.32	0.04	0.38	0.28	0.29
2007	5	8	200	193	39	0.07	0.38	0.17	0.51	0.41	0.40
2008	4	6	184	184	-	0.36	0.40	0.34	-	0.27	0.20
2009	5	6	190	177	47	0.22	0.24	0.21	0.34	0.28	0.26
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 269: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ciencias)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	6	8	319	115	299	0.06	0.14	-0.08	0.07	0.23	0.24
2005	6	10	394	87	385	0.09	0.18	0.12	0.11	0.31	0.31
2006	7	12	483	116	476	0.19	0.18	0.29	0.21	0.21	0.16
2007	8	15	583	137	575	0.03	0.14	-0.13	0.08	0.25	0.19
2008	8	16	556	127	534	0.19	0.17	0.45	0.17	0.23	0.23
2009	6	14	546	179	523	0.22	0.22	0.17	0.34	0.25	0.25
2010	2	2	20	-	20	0.19	-0.15	-	-0.36	0.34	-0.07

Table 270: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ciencias_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	18	549	130	549	0.09	0.24	-0.05	0.20	0.29	0.26
2005	11	23	818	187	816	0.07	0.22	0.01	0.22	0.18	0.18
2006	11	23	922	204	922	0.15	0.23	0.07	0.19	0.24	0.21
2007	12	24	968	193	968	0.18	0.18	0.58	0.19	0.22	0.14
2008	12	24	962	202	961	0.19	0.28	0.34	0.31	0.18	0.19
2009	12	24	864	188	860	0.16	0.29	0.17	0.21	0.20	0.11
2010	1	3	148	24	148	0.21	0.30	0.53	0.49	0.16	0.14

Table 271: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ciencias_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	5	137	48	137	0.06	0.17	0.07	0.14	0.22	0.27
2005	4	7	216	83	206	0.24	0.25	0.26	0.19	0.20	0.17
2006	6	10	259	94	250	0.16	0.12	-0.20	0.18	0.28	0.22
2007	6	8	223	69	217	0.36	0.41	0.12	0.30	0.33	0.27
2008	4	5	133	49	126	0.29	0.23	0.04	0.42	0.23	0.14
2009	4	5	112	-	112	0.11	0.32	-	0.23	0.22	0.27
2010	1	1	16	-	16	0.77	0.67	-	0.54	0.24	0.21

Table 272: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ciencias_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	39	-	39	0.16	0.51	-	0.46	0.41	-
2005	1	1	27	12	27	0.20	-0.17	-0.27	0.26	0.23	0.45
2006	2	2	46	-	46	0.44	0.43	-	0.07	0.51	0.31
2007	2	4	102	-	102	0.31	0.31	-	0.30	0.09	0.13
2008	2	4	133	47	133	0.12	0.45	0.22	0.19	0.14	0.11
2009	2	4	138	48	138	0.16	0.52	0.14	0.46	0.26	0.27
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 273: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ciencias_Sociales_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	14	21	689	651	273	0.13	0.25	0.02	0.17	0.37	0.31
2005	17	29	1208	1123	414	0.08	0.18	0.08	0.13	0.34	0.25
2006	17	28	1353	1216	515	0.12	0.21	0.13	0.18	0.38	0.28
2007	17	28	1386	1266	493	0.12	0.15	0.04	0.06	0.39	0.33
2008	16	27	1400	1269	498	0.09	0.13	0.10	-0.03	0.33	0.25
2009	16	28	1385	1238	512	0.15	0.23	0.09	0.19	0.39	0.32
2010	1	2	42	37	-	0.15	0.10	-0.34	-	0.66	0.56

Table 274: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ciencias_Sociales_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	6	11	370	370	89	0.05	0.18	0.03	0.21	0.31	0.36
2005	9	17	686	685	113	0.07	0.14	0.03	0.19	0.32	0.34
2006	12	20	853	852	151	0.05	0.13	-0.01	0.04	0.28	0.24
2007	13	27	1123	1117	185	0.11	0.11	0.07	0.17	0.31	0.26
2008	13	28	1182	1179	199	0.06	0.19	-0.03	0.40	0.32	0.26
2009	13	28	1336	1332	270	0.09	0.13	0.05	0.03	0.30	0.28
2010	1	2	89	89	-	-0.16	-0.18	0.07	-	0.24	0.18

Table 275: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ciencias_Sociales_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	11	247	247	62	0.12	0.15	0.04	0.11	0.30	0.31
2005	11	11	391	388	97	0.02	0.04	0.05	0.06	0.34	0.36
2006	12	13	545	543	122	0.04	0.11	-0.04	0.29	0.27	0.25
2007	12	14	572	569	135	-0.03	0.30	0.07	-0.14	0.24	0.17
2008	11	13	523	521	126	0.01	0.02	-0.05	0.10	0.21	0.14
2009	12	14	590	588	178	0.01	0.19	0.04	-0.07	0.25	0.19
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 276: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Comunicaciones)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	22	21	9	-0.06	-0.12	-0.02	-0.40	0.41	0.50
2005	1	1	33	31	-	0.35	0.12	0.41	-	0.26	0.51
2006	2	2	60	57	22	-0.07	-0.06	-0.10	-0.50	0.16	0.39
2007	2	2	69	65	-	0.10	0.02	0.00	-	0.03	0.05
2008	2	2	62	59	-	-0.20	0.02	-0.11	-	0.18	0.15
2009	1	1	38	36	-	0.10	-0.34	-0.01	-	0.12	0.31
2010	1	1	23	23	-	0.32	0.48	-0.08	-	0.48	0.53

Table 277: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Construcción)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	10	378	159	373	0.14	0.11	0.14	0.16	0.20	0.23
2005	11	11	456	195	447	0.16	0.16	0.17	0.10	0.24	0.20
2006	12	12	570	236	554	0.07	0.20	0.09	0.11	0.17	0.20
2007	12	12	652	250	635	0.09	0.23	-0.01	0.16	0.22	0.27
2008	13	13	664	272	646	0.03	0.13	0.00	0.10	0.18	0.20
2009	13	13	692	268	672	0.00	0.28	0.02	0.23	0.17	0.14
2010	1	1	1	-	-	-	-	-	-	-	0.49

Table 278: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Derecho)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	15	18	1064	1060	279	0.12	0.16	0.06	0.04	0.35	0.27
2005	14	18	1394	1391	303	0.11	0.13	0.09	0.16	0.35	0.30
2006	15	19	1655	1646	350	0.15	0.16	0.10	0.15	0.34	0.27
2007	15	19	1746	1737	313	0.14	0.16	0.05	0.11	0.34	0.28
2008	14	18	1817	1814	313	0.09	0.19	0.09	0.11	0.33	0.29
2009	14	19	1707	1692	316	0.05	0.26	0.15	0.08	0.32	0.34
2010	2	3	337	337	73	0.13	0.28	0.02	-0.21	0.34	0.29

Table 279: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Diseño)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	8	10	392	386	77	0.15	0.19	0.11	-0.10	0.29	0.28
2005	11	15	644	618	152	-0.03	0.19	0.00	0.03	0.36	0.30
2006	12	18	861	807	247	0.05	0.17	0.07	0.04	0.25	0.24
2007	11	17	857	792	250	-0.07	0.09	-0.06	0.04	0.26	0.18
2008	11	15	826	754	235	0.09	0.23	0.10	0.04	0.35	0.32
2009	11	17	745	691	211	0.09	0.16	0.09	0.26	0.29	0.28
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 280: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Educación)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	13	26	513	513	148	-0.04	0.09	0.00	0.08	0.33	0.35
2005	16	49	1292	1284	310	0.02	0.17	-0.01	0.27	0.32	0.27
2006	17	53	1586	1581	330	0.08	0.10	-0.02	-0.20	0.32	0.28
2007	17	57	1797	1778	345	0.05	0.13	-0.01	-0.09	0.31	0.26
2008	17	55	1741	1723	376	0.02	0.14	0.04	0.14	0.31	0.26
2009	17	56	1726	1694	418	0.08	0.20	0.04	0.01	0.27	0.21
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 281: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Educación_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	12	24	579	541	219	-0.02	0.21	0.02	0.07	0.36	0.30
2005	14	45	1405	1308	560	0.01	0.17	0.03	0.09	0.26	0.27
2006	14	48	1737	1578	742	0.08	0.09	0.03	0.06	0.28	0.26
2007	16	52	1942	1743	793	0.04	0.09	0.04	0.09	0.28	0.26
2008	15	49	1912	1708	831	0.03	0.12	0.05	0.17	0.25	0.22
2009	15	48	1983	1801	848	0.06	0.24	0.05	0.12	0.28	0.21
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 282: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Educación_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	9	11	232	111	230	0.10	0.24	0.22	0.13	0.27	0.28
2005	11	27	768	303	767	0.09	0.20	0.10	0.21	0.27	0.21
2006	12	32	957	397	947	0.20	0.21	0.20	0.19	0.26	0.21
2007	11	28	892	365	884	0.01	0.10	-0.03	0.00	0.21	0.15
2008	12	27	844	291	837	-0.02	0.12	0.04	0.12	0.33	0.27
2009	12	31	910	341	905	0.13	0.28	0.18	0.20	0.31	0.27
2010	1	1	20	-	20	-0.07	-0.47	-	0.24	0.54	0.06

Table 283: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Educación_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	8	10	231	129	189	0.13	0.14	0.05	0.09	0.27	0.29
2005	8	13	452	281	357	-0.05	0.11	0.01	0.10	0.29	0.26
2006	11	17	634	400	509	0.17	0.18	-0.03	0.04	0.32	0.27
2007	11	17	643	363	506	0.12	0.10	0.05	0.10	0.25	0.17
2008	11	17	661	346	535	0.04	0.08	0.01	0.10	0.26	0.21
2009	11	17	692	364	568	-0.07	0.13	-0.14	0.14	0.27	0.22
2010	1	1	7	-	-	-	-	-	-	-0.19	0.15

Table 284: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (General)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	183	73	156	0.02	0.03	0.01	0.34	0.27	0.21
2005	1	1	222	86	202	0.02	-0.14	0.06	0.22	0.24	0.23
2006	1	1	238	87	223	-0.10	0.00	-0.33	-0.02	0.33	0.26
2007	1	1	248	78	229	0.17	-0.12	-0.06	0.01	0.16	0.08
2008	1	1	248	81	234	0.11	-0.04	0.36	0.13	0.16	0.15
2009	1	1	192	68	171	-0.01	-0.06	-0.07	-0.12	0.10	0.15
2010	1	1	237	102	199	0.20	0.00	0.26	0.21	0.33	0.17

Table 285: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Humanidades)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	5	8	248	248	49	0.09	0.16	0.10	0.26	0.34	0.21
2005	8	11	428	425	-	0.06	0.10	0.00	-	0.34	0.31
2006	8	13	449	445	87	0.18	0.16	0.17	0.17	0.28	0.35
2007	8	13	495	489	72	0.10	0.26	0.05	0.54	0.40	0.25
2008	9	14	453	450	58	0.18	0.15	0.11	0.79	0.27	0.18
2009	9	13	448	443	-	-0.04	0.16	0.08	-	0.30	0.26
2010	1	3	129	129	-	0.20	0.06	0.04	-	0.28	0.18

Table 286: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ingeniería_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	21	78	3370	1054	3354	0.15	0.18	0.08	0.21	0.32	0.29
2005	21	91	4687	1563	4624	0.19	0.16	0.12	0.24	0.28	0.26
2006	21	95	5089	1807	4997	0.15	0.17	0.02	0.20	0.28	0.23
2007	21	107	5709	1983	5572	0.13	0.20	0.08	0.19	0.27	0.23
2008	22	109	5735	1865	5611	0.17	0.19	0.07	0.26	0.26	0.23
2009	22	101	5813	1923	5687	0.18	0.23	0.08	0.26	0.28	0.24
2010	1	1	128	36	128	0.22	0.01	0.56	0.35	-0.06	-0.06

Table 287: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ingeniería_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	9	23	515	157	511	0.05	0.09	-0.02	0.07	0.18	0.21
2005	13	39	1062	447	995	0.14	0.08	0.10	0.10	0.24	0.19
2006	15	48	1348	522	1257	0.11	0.13	0.05	0.15	0.23	0.15
2007	13	47	1444	518	1331	0.09	0.21	0.16	0.13	0.21	0.17
2008	15	52	1499	564	1354	0.11	0.10	-0.01	0.10	0.26	0.22
2009	15	56	1544	583	1430	0.13	0.20	0.22	0.14	0.25	0.21
2010	1	1	27	-	27	-0.18	0.09	-	-0.12	0.43	-0.09

Table 288: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Ingeniería_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	5	115	35	115	0.08	0.28	0.02	0.25	0.28	0.28
2005	4	6	175	-	175	0.23	0.14	-	0.25	0.29	0.22
2006	4	7	257	71	256	0.19	0.15	0.11	0.28	0.30	0.31
2007	6	12	421	121	420	0.24	0.21	-0.32	0.21	0.29	0.22
2008	6	11	387	83	386	0.20	0.14	0.14	0.23	0.33	0.30
2009	6	12	366	91	366	0.11	0.18	-0.25	0.16	0.24	0.26
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 289: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Mar)

`	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	5	6	113	58	113	0.20	0.47	0.15	0.23	0.22	0.13
2005	8	11	234	81	229	0.29	0.31	0.60	0.27	0.35	0.30
2006	7	7	191	79	188	0.14	0.24	0.00	0.03	0.24	0.05
2007	4	5	152	-	152	0.21	0.37	-	0.36	0.22	0.17
2008	8	10	244	71	239	0.17	0.31	-0.29	0.22	0.35	0.31
2009	5	7	142	-	140	0.39	0.38	-	0.34	0.40	0.33
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 290: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Periodismo)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	10	404	402	82	0.14	0.17	0.13	0.39	0.40	0.32
2005	12	12	543	543	90	0.00	0.14	0.02	-0.30	0.50	0.44
2006	13	14	672	670	104	0.10	0.11	0.06	-0.13	0.43	0.36
2007	14	14	687	686	104	0.08	0.23	0.00	0.21	0.34	0.34
2008	13	13	558	557	92	0.01	0.09	0.11	0.78	0.35	0.24
2009	12	12	548	546	86	0.10	0.16	0.00	-0.24	0.37	0.31
2010	1	1	46	46	-	0.20	-0.04	0.12	-	0.43	0.60

Table 291: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Salud_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	13	21	895	201	894	0.04	-0.02	0.09	0.18	0.32	0.23
2005	15	27	1476	336	1475	0.06	0.03	0.01	0.22	0.28	0.27
2006	14	26	1721	394	1719	0.16	0.09	-0.02	0.17	0.22	0.22
2007	14	26	1774	412	1771	0.09	0.06	0.06	0.07	0.24	0.20
2008	15	26	1657	358	1654	0.00	0.07	0.04	0.15	0.25	0.23
2009	14	27	1797	412	1794	0.04	0.08	0.05	0.16	0.32	0.28
2010	2	4	68	-	68	0.00	0.33	-	-0.22	0.46	0.04

Table 292: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Salud_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	20	41	1255	349	1251	0.05	0.10	0.15	0.17	0.24	0.28
2005	21	49	2148	615	2141	0.13	0.11	0.01	0.18	0.28	0.27
2006	20	48	2466	649	2453	0.15	0.11	0.17	0.19	0.28	0.25
2007	20	50	2573	631	2557	0.12	0.13	0.08	0.15	0.24	0.21
2008	20	49	2561	636	2547	0.06	0.13	0.10	0.16	0.21	0.18
2009	20	50	2484	600	2469	0.10	0.13	0.02	0.14	0.23	0.20
2010	2	6	219	-	219	0.04	-0.07	-	0.02	0.20	0.16

Table 293: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Salud_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	17	517	146	511	0.08	0.23	0.11	0.18	0.24	0.26
2005	14	23	998	276	995	0.10	0.11	0.09	0.20	0.31	0.30
2006	14	23	1121	296	1115	0.13	0.19	0.08	0.18	0.27	0.25
2007	15	24	1197	284	1191	0.11	0.15	0.08	0.24	0.24	0.20
2008	15	24	1152	252	1150	0.15	0.14	0.34	0.18	0.27	0.23
2009	15	23	993	235	987	0.17	0.20	0.13	0.15	0.27	0.24
2010	2	3	111	24	111	-0.24	0.23	-0.37	0.03	0.44	0.24

Table 294: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Técnico_Administración)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	87	80	48	0.05	0.20	0.08	0.09	0.22	0.45
2005	5	6	201	180	63	0.01	0.00	-0.06	0.27	0.41	0.34
2006	5	6	257	212	99	0.03	0.07	0.03	-0.01	0.31	0.30
2007	5	7	249	217	88	0.17	0.22	0.11	0.25	0.27	0.16
2008	6	8	287	244	106	0.13	0.10	0.21	0.23	0.26	0.25
2009	6	7	236	195	98	0.05	0.03	-0.05	-0.11	0.18	0.13
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 295: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Técnico_Agro)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	24	15	24	0.20	0.21	0.28	0.13	-0.06	-
2005	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2006	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2007	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2008	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2009	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 296: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Técnico_Ciencias)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	24	-	24	0.03	0.00	-	-0.10	0.18	-0.27
2005	2	3	74	30	72	0.34	-0.16	0.03	0.32	0.36	0.52
2006	2	3	86	-	85	0.03	-0.08	-	0.12	0.32	0.35
2007	2	3	70	34	69	0.29	0.36	0.39	0.50	0.25	0.18
2008	3	4	116	41	114	0.12	0.16	-0.37	0.06	0.30	0.28
2009	3	4	109	-	106	-0.07	0.28	-	0.37	0.40	0.43
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 297: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Técnico_Diseño)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	15	15	-	-0.07	0.37	0.09	-	0.32	-0.03
2005	2	2	41	37	-	0.15	-0.02	0.46	-	-0.14	0.08
2006	3	3	62	49	36	-0.01	0.23	-0.05	-0.15	0.04	0.17
2007	3	3	62	46	-	0.07	0.36	-0.06	-	0.19	0.10
2008	2	2	49	32	27	0.05	0.38	-0.30	0.65	0.21	0.15
2009	2	2	48	-	35	0.06	0.15	-	0.14	-0.13	-0.03
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 298: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Técnico_Educación)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	22	15	19	-0.52	-0.36	0.02	0.48	0.57	-
2005	1	1	19	-	18	-0.25	0.28	-	0.15	0.09	0.11
2006	1	1	18	11	15	0.33	-0.15	-0.20	0.08	0.10	-0.20
2007	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2008	1	1	15	-	-	-0.24	0.19	-	-	0.53	0.46
2009	1	1	21	-	16	0.48	-0.11	-	0.41	-0.16	0.05
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 299: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Técnico_Idioma)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	59	59	14	-0.03	0.30	0.10	0.09	0.30	0.39
2005	3	3	111	108	42	0.19	0.19	0.13	0.04	0.37	0.25
2006	3	3	122	119	41	-0.14	0.04	-0.07	-0.04	0.30	0.49
2007	3	3	125	119	44	0.24	0.19	-0.11	0.44	0.33	0.13
2008	3	3	129	125	39	0.09	0.39	0.19	0.01	0.26	0.25
2009	3	3	114	106	32	0.09	0.10	0.10	-0.07	0.10	0.21
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 300: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Técnico_Ingeniería)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	15	364	135	346	0.06	0.19	0.10	0.13	0.25	0.22
2005	4	39	1071	493	960	0.19	0.22	0.08	0.18	0.22	0.22
2006	6	45	1286	606	1133	0.09	0.19	0.08	0.11	0.21	0.19
2007	5	46	1360	606	1177	0.13	0.16	0.03	0.15	0.21	0.17
2008	6	46	1631	750	1367	0.08	0.15	0.06	0.12	0.25	0.23
2009	6	48	1585	718	1367	0.02	0.15	0.10	0.11	0.19	0.18
2010	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 301: Average Predictive Validity Coefficient (Corrected by Range Restrictions) of Predictor Measures on University SYGPA by Admission Year (Veterinaria)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	5	240	66	240	0.14	0.16	0.00	0.15	0.30	0.28
2005	4	5	367	86	366	0.16	0.16	0.30	0.20	0.22	0.14
2006	4	5	381	92	381	0.09	0.11	0.09	0.14	0.19	0.18
2007	4	5	417	103	414	0.05	0.11	-0.02	0.03	0.31	0.27
2008	4	5	379	92	377	0.15	0.14	-0.10	0.18	0.22	0.19
2009	4	5	355	73	352	-0.04	0.14	0.10	0.10	0.41	0.30
2010	1	1	100	21	100	0.22	0.37	0.06	0.52	0.12	0.33

Appendix R. Prediction Validity by the Type of Career - University Completion

Table 302: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Administración)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	20	26	1353	1240	741	0.04	0.06	0.08	0.06	0.07	0.10
2005	19	27	1967	1738	990	0.09	0.13	0.08	0.01	0.15	0.17
2006	19	27	2249	1837	1123	-0.01	-0.02	0.03	-0.01	0.03	0.04

Table 303: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Administración_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	12	14	477	450	215	-0.01	0.08	0.00	-0.01	0.04	0.08
2005	13	15	756	673	290	0.06	0.08	0.03	0.14	0.14	0.16
2006	12	16	864	740	376	0.04	0.09	-0.06	0.14	0.06	0.10

Table 304: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Administración_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	62	57	-	0.03	-0.16	-0.04	-	0.21	0.06
2005	4	5	189	162	45	0.01	0.12	0.00	0.03	0.03	0.07
2006	3	5	215	191	75	-0.08	-0.04	0.03	0.12	0.06	0.11

Table 305: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Agro)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	8	12	568	223	567	0.08	0.11	0.11	0.07	0.05	0.08
2005	7	11	737	242	732	0.11	0.05	0.01	0.12	0.12	0.14
2006	7	10	808	273	798	0.03	0.11	-0.04	0.07	0.07	0.08

Table 306: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Agro_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	2	2	54	32	47	0.07	0.05	0.23	-0.06	-0.12	0.22
2005	3	4	112	72	88	-0.01	0.22	0.07	0.06	0.08	0.01
2006	3	3	92	58	69	0.13	0.21	0.20	0.02	0.07	-0.04

Table 307: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Arquitectura)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	16	16	863	833	393	0.02	0.14	0.05	0.10	0.12	0.10
2005	18	18	1194	1114	529	0.01	0.02	0.01	0.06	0.01	0.03
2006	12	12	957	848	433	-0.01	-0.03	0.00	-0.04	-0.02	-0.04

Table 308: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Arte_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	5	199	196	36	0.06	0.21	-0.05	0.06	0.14	0.04
2005	5	7	317	317	45	0.08	0.02	0.05	-0.73	0.15	0.16
2006	6	7	350	345	72	0.12	0.25	0.00	0.34	0.07	0.04

Table 309: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Arte_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	4	102	102	13	0.00	0.01	0.04	0.10	0.19	0.11
2005	5	9	256	245	62	0.02	0.15	0.06	0.03	0.16	0.13
2006	4	7	228	219	48	-0.04	0.02	-0.16	-0.08	0.08	0.10

Table 310: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ciencias)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	6	9	406	140	380	0.05	0.05	0.05	0.07	0.02	0.04
2005	6	10	503	117	482	-0.12	0.06	-0.15	0.09	0.10	0.08
2006	7	12	624	145	594	-0.04	0.06	-0.02	0.01	0.05	0.00

Table 311: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ciencias_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	21	637	156	637	-0.02	0.17	0.02	0.06	0.19	0.19
2005	9	20	848	195	843	0.02	0.10	0.13	0.13	0.08	0.08
2006	10	19	840	185	837	-0.08	0.02	-0.12	0.05	0.07	0.06

Table 312: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ciencias_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	7	223	91	217	0.11	0.20	0.23	0.13	0.02	0.09
2005	7	10	360	144	340	0.14	0.10	0.03	0.21	0.07	0.17
2006	6	11	387	148	362	-0.06	0.09	0.00	0.14	0.00	0.02

Table 313: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ciencias_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	48	-	48	-0.10	0.10	-	0.11	0.23	0.18
2005	2	2	85	39	85	0.14	0.17	0.20	-0.02	0.04	-0.13
2006	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 314: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ciencias_Sociales_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	17	23	872	825	346	-0.05	0.05	0.03	-0.03	0.15	0.13
2005	18	30	1433	1323	464	0.08	0.10	0.08	-0.09	0.16	0.09
2006	14	22	1272	1071	458	0.03	0.02	-0.02	-0.12	0.12	0.03

Table 315: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ciencias_Sociales_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	6	12	508	508	115	-0.03	0.09	0.02	0.00	0.09	0.03
2005	10	16	814	812	134	-0.06	0.00	-0.02	0.02	0.12	0.09
2006	11	18	983	943	168	0.03	0.06	0.02	0.00	0.08	0.09

Table 316: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ciencias_Sociales_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	11	304	303	74	0.01	-0.07	-0.02	0.24	0.13	0.08
2005	11	12	460	449	94	-0.04	0.05	0.04	-0.11	0.06	0.14
2006	12	13	641	629	137	0.04	0.06	0.03	-0.12	0.08	0.18

Table 317: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Comunicaciones)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	32	31	-	-0.01	0.17	0.11	-	0.09	-0.09
2005	1	1	41	39	-	0.03	-0.08	0.04	-	-0.03	0.04
2006	2	2	77	71	31	-0.17	-0.04	-0.16	-0.35	0.37	0.01

Table 318: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Construcción)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	11	11	645	301	621	0.02	0.06	0.07	0.02	0.05	0.08
2005	12	12	817	373	774	0.06	0.08	0.11	0.02	0.02	0.01
2006	10	10	878	386	828	0.03	-0.02	-0.04	-0.10	0.05	0.07

Table 319: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Derecho)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	17	20	1462	1458	394	0.05	0.02	0.01	0.07	0.07	0.04
2005	17	21	2014	1990	450	0.02	0.02	-0.04	0.05	0.07	0.06
2006	10	12	1322	1291	274	0.02	-0.05	-0.06	0.13	0.04	0.00

Table 320: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Diseño)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	10	12	530	515	130	-0.07	0.04	0.08	-0.02	0.03	0.00
2005	12	15	673	628	184	0.03	0.07	-0.02	0.02	0.14	0.15
2006	12	15	919	817	256	-0.04	0.10	0.02	0.00	0.13	0.10

Table 321: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Educación)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	14	26	546	546	145	0.00	0.07	-0.03	0.08	0.11	0.11
2005	17	52	1561	1540	337	-0.02	0.08	0.00	0.16	0.14	0.15
2006	14	49	1669	1625	308	-0.03	0.01	-0.02	-0.22	0.12	0.12

Table 322: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Educación_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	13	22	541	505	193	-0.04	0.06	-0.04	-0.07	0.09	0.07
2005	13	43	1427	1285	544	-0.09	0.04	-0.06	0.04	0.04	0.03
2006	13	46	1672	1457	620	0.03	0.09	-0.01	-0.02	0.13	0.07

Table 323: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Educación_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	9	15	486	206	486	0.03	0.12	0.06	0.07	0.10	0.09
2005	10	28	885	326	882	0.00	0.11	-0.07	0.09	0.10	0.09
2006	10	30	1065	418	1052	-0.02	0.07	0.04	0.04	0.01	0.02

Table 324: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Educación_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	9	11	279	163	225	-0.02	0.12	-0.04	-0.01	0.12	0.07
2005	8	13	508	311	385	0.01	0.01	0.05	-0.07	0.12	0.20
2006	10	15	685	408	497	0.03	0.09	0.09	0.11	0.08	0.09

Table 325: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (General)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	196	78	166	-0.16	0.04	0.00	0.03	0.19	0.11
2005	1	1	249	92	229	0.09	0.05	-0.06	-0.06	0.01	-0.05
2006	1	1	259	90	244	-0.07	0.07	0.04	0.08	-0.01	-0.01

Table 326: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Humanidades)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	5	8	277	276	58	0.07	-0.01	0.06	-0.67	0.07	-0.08
2005	9	13	542	524	-	0.09	-0.03	0.02	-	0.11	0.07
2006	7	11	536	526	97	0.06	0.00	0.09	0.01	0.08	0.13

Table 327: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ingeniería_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	21	80	4024	1266	4003	0.04	0.03	0.03	0.04	0.06	0.06
2005	22	87	5486	1836	5292	0.00	-0.01	-0.05	0.02	0.02	0.03
2006	21	83	5598	1913	5242	-0.04	-0.03	-0.03	-0.03	-0.03	-0.04

Table 328: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ingeniería_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	15	36	942	320	919	-0.02	0.05	-0.05	0.01	0.01	0.05
2005	13	47	1607	643	1462	0.04	0.04	0.07	0.02	0.03	0.05
2006	13	53	1869	688	1672	-0.02	0.00	-0.02	-0.02	0.00	0.05

Table 329: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Ingeniería_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	3	5	130	39	130	0.06	0.08	-0.11	0.02	0.27	0.21
2005	3	5	170	-	168	0.03	0.12	-	0.06	0.09	0.02
2006	4	8	316	78	306	-0.02	0.05	0.39	0.03	0.06	0.08

Table 330: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Mar)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	7	8	158	56	155	-0.07	-0.02	-0.18	-0.07	0.05	0.13
2005	7	9	298	103	292	0.01	0.02	0.03	0.09	0.12	0.11
2006	6	7	240	103	230	0.12	0.09	0.02	-0.02	0.02	0.02

Table 331: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Periodismo)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	13	13	505	503	95	-0.01	0.05	0.00	-0.03	0.13	0.13
2005	11	11	548	536	101	-0.05	0.00	0.03	-0.05	0.15	0.13
2006	12	13	648	610	99	-0.10	-0.04	0.03	-0.25	0.15	0.14

Table 332: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Salud_1)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	13	20	932	213	930	0.02	0.00	-0.07	0.07	0.15	0.17
2005	11	19	1174	263	1169	-0.03	0.01	-0.01	0.08	0.10	0.03
2006	10	18	1435	315	1399	-0.01	-0.10	-0.03	-0.07	-0.03	-0.04

Table 333: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Salud_2)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	15	37	1257	336	1253	0.00	0.09	-0.06	0.10	0.06	0.08
2005	18	46	2321	626	2262	0.10	0.12	0.02	0.14	0.11	0.15
2006	18	46	2659	677	2552	0.10	0.10	0.12	0.17	0.13	0.11

Table 334: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Salud_3)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	7	14	468	128	461	0.13	0.19	0.04	0.15	0.08	0.10
2005	12	20	1029	263	1006	0.08	0.16	0.12	0.15	0.06	0.14
2006	12	20	1142	269	1097	0.05	0.18	-0.05	0.12	0.17	0.13

Table 335: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Técnico_Administración)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	4	142	133	65	0.00	-0.05	0.06	0.03	0.10	0.21
2005	5	6	265	233	92	0.07	0.07	0.07	-0.09	0.04	-0.02
2006	5	6	328	266	118	-0.04	0.03	-0.05	0.14	0.17	0.14

Table 336: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Técnico_Agro)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	20	14	20	0.11	0.16	0.12	-0.02	0.28	0.23
2005	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2006	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 337: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Técnico_Ciencias)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	27	-	27	-0.20	0.16	-	-0.13	0.14	0.23
2005	2	3	79	30	76	0.03	-0.02	-0.11	0.03	0.02	0.03
2006	2	3	94	-	92	0.13	0.12	-	0.13	-0.04	0.14

Table 338: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Técnico_Diseño)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	17	17	-	-0.45	0.22	-0.24	-	0.33	-0.10
2005	2	2	51	46	-	-0.12	-0.21	-0.14	-	0.09	0.00
2006	3	3	74	58	42	-0.21	0.24	-0.24	-0.12	-0.01	0.18

Table 339: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Técnico_Educación)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	1	1	31	20	26	-0.37	-0.06	-0.32	0.40	0.28	-
2005	1	1	22	13	19	-0.12	0.42	-0.28	-0.11	0.47	0.41
2006	1	1	27	17	22	0.19	-0.15	0.00	0.14	-0.01	0.04

Table 340: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Técnico_Idioma)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	4	174	172	57	0.04	0.03	0.07	0.06	0.03	-0.10
2005	4	6	280	268	100	0.05	0.20	0.02	0.08	0.05	0.03
2006	5	7	330	311	111	0.15	0.01	0.12	0.07	0.05	0.04

Table 341: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Técnico_Ingeniería)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	9	22	545	197	528	0.00	0.09	0.08	0.07	0.05	0.08
2005	6	41	1287	563	1156	-0.01	0.07	-0.03	0.07	0.03	0.05
2006	5	44	1402	636	1229	0.01	0.08	-0.10	0.05	-0.05	0.00

Table 342: Average Pearson Correlations (Corrected by Range Restrictions) of Predictor Measures on University Completion by Admission Year (Veterinaria)

	Sample Sizes					PSU Tests				High School	
			Student								
Year	University	Career	L&M	H	S	Language	Mathematics	History	Science	NEM	Rank
2004	4	5	269	74	269	0.11	0.11	0.19	0.18	0.13	0.22
2005	3	4	345	83	338	0.16	0.06	0.20	0.26	0.05	0.07
2006	3	4	355	88	345	0.06	0.06	0.16	0.05	0.13	0.13

Appendix S. Incremental Prediction Validity of Ranking by the Type of Career – First Year Grade Point Average (FYGPA)

Table 343: Average R-square for Base and Revised Models and FYGPA by Admission Year (Administración)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	18	23	1158	0.15	0.16	0.01
2005	18	27	1734	0.20	0.20	0.01
2006	20	30	2092	0.16	0.17	0.01
2007	20	30	2487	0.15	0.15	0.00
2008	20	29	2873	0.14	0.15	0.00
2009	21	30	3033	0.12	0.13	0.00
2010	19	27	2942	0.06	0.07	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 344: Average R-square for Base and Revised Models and FYGPA by Admission Year (Administración_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	12	371	0.15	0.17	0.02
2005	12	14	564	0.21	0.25	0.03
2006	15	19	898	0.11	0.12	0.01
2007	15	19	1118	0.13	0.15	0.02
2008	15	18	1008	0.10	0.11	0.01
2009	16	19	1078	0.13	0.14	0.01
2010	14	16	991	0.10	0.11	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 345: Average R-square for Base and Revised Models and FYGPA by Admission Year (Administración_2)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	4	96	0.17	0.17	0.00
2005	4	5	164	0.23	0.23	0.00
2006	4	7	284	0.22	0.22	0.01
2007	4	7	296	0.24	0.25	0.00
2008	6	9	394	0.12	0.12	0.00
2009	6	10	392	0.15	0.16	0.00
2010	6	10	428	0.09	0.09	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 346: Average R-square for Base and Revised Models and FYGPA by Admission Year (Agro)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	8	11	516	0.17	0.17	0.00
2005	8	13	796	0.20	0.20	0.00
2006	8	13	924	0.22	0.23	0.01
2007	8	14	991	0.19	0.20	0.00
2008	8	14	972	0.15	0.16	0.00
2009	8	14	948	0.19	0.19	0.00
2010	8	13	917	0.15	0.15	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 347: Average R-square for Base and Revised Models and FYGPA by Admission Year (Agro_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	2	2	51	0.00	0.23	0.22
2005	3	4	99	0.24	0.24	0.00
2006	3	3	84	0.10	0.10	0.00
2007	3	4	109	0.46	0.47	0.02
2008	3	4	95	0.22	0.23	0.00
2009	3	4	116	0.35	0.36	0.01
2010	3	4	94	0.24	0.26	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 348: Average R-square for Base and Revised Models and FYGPA by Admission Year (Arquitectura)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	11	580	0.14	0.14	0.00
2005	12	12	789	0.19	0.19	0.00
2006	12	12	887	0.17	0.17	0.00
2007	12	12	892	0.21	0.21	0.00
2008	14	15	952	0.19	0.19	0.00
2009	14	15	993	0.16	0.16	0.00
2010	12	13	731	0.15	0.16	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 349: Average R-square for Base and Revised Models and FYGPA by Admission Year (Arte_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	4	170	0.10	0.11	0.01
2005	6	8	298	0.13	0.14	0.01
2006	8	9	377	0.08	0.08	0.00
2007	9	11	406	0.10	0.10	0.01
2008	7	8	356	0.15	0.16	0.01
2009	8	10	418	0.12	0.12	0.00
2010	8	11	382	0.09	0.11	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 350: Average R-square for Base and Revised Models and FYGPA by Admission Year (Arte_2)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	4	96	0.38	0.38	0.00
2005	5	9	233	0.12	0.13	0.01
2006	5	7	222	0.03	0.03	0.00
2007	5	9	242	0.05	0.06	0.00
2008	5	9	262	0.05	0.06	0.01
2009	6	10	292	0.06	0.07	0.01
2010	5	7	156	0.18	0.19	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 351: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ciencias)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	7	9	330	0.17	0.17	0.00
2005	7	12	473	0.19	0.19	0.00
2006	8	14	553	0.15	0.17	0.02
2007	8	15	666	0.14	0.14	0.00
2008	8	16	807	0.13	0.13	0.00
2009	7	15	939	0.08	0.10	0.02
2010	7	15	945	0.09	0.09	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 352: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ciencias_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	20	607	0.20	0.21	0.01
2005	10	22	855	0.20	0.20	0.00
2006	11	23	973	0.23	0.23	0.00
2007	12	24	1059	0.18	0.19	0.01
2008	12	24	1086	0.20	0.20	0.00
2009	12	25	1074	0.17	0.17	0.00
2010	11	22	986	0.15	0.15	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 353: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ciencias_2)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	6	169	0.24	0.24	0.00
2005	5	8	277	0.24	0.24	0.00
2006	6	10	322	0.29	0.30	0.01
2007	6	10	349	0.22	0.22	0.00
2008	5	9	277	0.17	0.17	0.00
2009	5	8	246	0.15	0.15	0.00
2010	5	9	243	0.25	0.25	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 354: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ciencias_3)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	17	0.03	0.04	0.01
2005	2	2	54	0.31	0.32	0.01
2006	2	2	46	0.15	0.16	0.00
2007	2	4	104	0.06	0.06	0.00
2008	2	4	143	0.05	0.08	0.02
2009	2	4	153	0.18	0.18	0.00
2010	2	4	158	0.09	0.10	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 355: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ciencias_Sociales_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	13	20	718	0.21	0.23	0.01
2005	15	26	1180	0.11	0.11	0.00
2006	18	29	1471	0.25	0.25	0.00
2007	17	28	1509	0.18	0.18	0.01
2008	18	29	1551	0.18	0.18	0.00
2009	18	30	1586	0.13	0.13	0.00
2010	17	29	1513	0.17	0.17	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 356: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ciencias_Sociales_2)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	6	13	490	0.05	0.06	0.01
2005	10	18	800	0.16	0.17	0.01
2006	12	20	1000	0.10	0.11	0.01
2007	13	28	1314	0.13	0.13	0.00
2008	13	29	1386	0.11	0.11	0.00
2009	13	29	1520	0.07	0.08	0.00
2010	12	27	1368	0.09	0.09	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 357: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ciencias_Sociales_3)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	11	260	0.18	0.22	0.03
2005	10	10	401	0.11	0.12	0.02
2006	12	13	571	0.11	0.11	0.00
2007	12	14	610	0.05	0.06	0.01
2008	11	13	570	0.11	0.12	0.00
2009	12	14	636	0.07	0.07	0.00
2010	11	12	558	0.06	0.06	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 358: Average R-square for Base and Revised Models and FYGPA by Admission Year (Comunicaciones)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	23	0.11	0.28	0.17
2005	1	1	36	0.09	0.15	0.06
2006	2	2	72	0.06	0.07	0.01
2007	2	2	79	0.10	0.13	0.03
2008	2	2	68	0.01	0.06	0.05
2009	2	2	54	0.19	0.19	0.00
2010	2	2	72	0.03	0.04	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 359: Average R-square for Base and Revised Models and FYGPA by Admission Year (Construcción)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	10	375	0.13	0.13	0.00
2005	13	13	535	0.07	0.07	0.00
2006	13	13	657	0.08	0.08	0.00
2007	14	14	775	0.14	0.15	0.02
2008	14	14	800	0.08	0.09	0.00
2009	14	14	844	0.08	0.08	0.00
2010	11	11	690	0.05	0.05	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 360: Average R-square for Base and Revised Models and FYGPA by Admission Year (Derecho)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	15	18	1175	0.19	0.20	0.01
2005	15	19	1548	0.16	0.16	0.00
2006	15	19	1756	0.18	0.18	0.00
2007	15	19	1941	0.17	0.17	0.00
2008	15	19	2053	0.14	0.14	0.00
2009	14	19	2038	0.15	0.16	0.00
2010	14	19	2014	0.17	0.17	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 361: Average R-square for Base and Revised Models and FYGPA by Admission Year (Diseño)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	9	11	434	0.02	0.02	0.00
2005	11	15	692	0.04	0.05	0.00
2006	12	18	932	0.08	0.08	0.00
2007	11	17	956	0.02	0.02	0.00
2008	12	17	961	0.06	0.06	0.00
2009	12	18	941	0.08	0.08	0.00
2010	11	15	620	0.07	0.07	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 362: Average R-square for Base and Revised Models and FYGPA by Admission Year (Educación)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	13	27	559	0.10	0.11	0.01
2005	15	49	1372	0.12	0.13	0.00
2006	16	55	1749	0.08	0.08	0.00
2007	16	59	2000	0.08	0.08	0.00
2008	16	57	1953	0.08	0.08	0.00
2009	17	57	1948	0.09	0.09	0.00
2010	16	53	1798	0.10	0.10	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 363: Average R-square for Base and Revised Models and FYGPA by Admission Year (Educación_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	12	27	643	0.08	0.08	0.00
2005	13	41	1404	0.08	0.09	0.01
2006	14	49	1830	0.06	0.06	0.00
2007	16	51	2032	0.10	0.10	0.00
2008	16	50	2046	0.06	0.06	0.00
2009	15	47	2115	0.06	0.07	0.00
2010	15	47	1890	0.03	0.03	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 364: Average R-square for Base and Revised Models and FYGPA by Admission Year (Educación_2)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	14	303	0.09	0.09	0.00
2005	11	26	803	0.11	0.11	0.00
2006	13	33	1063	0.11	0.11	0.00
2007	12	32	1097	0.10	0.10	0.00
2008	12	28	1004	0.09	0.09	0.00
2009	12	32	1114	0.10	0.10	0.00
2010	11	28	953	0.05	0.05	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 365: Average R-square for Base and Revised Models and FYGPA by Admission Year (Educación_3)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	8	10	247	0.07	0.07	0.00
2005	8	13	474	0.07	0.07	0.00
2006	11	17	675	0.14	0.14	0.00
2007	11	17	691	0.18	0.18	0.00
2008	11	17	710	0.21	0.21	0.00
2009	11	17	754	0.21	0.21	0.00
2010	10	16	719	0.12	0.13	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 366: Average R-square for Base and Revised Models and FYGPA by Admission Year (General)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	183	0.04	0.05	0.01
2005	1	1	233	0.02	0.02	0.00
2006	1	1	242	0.04	0.05	0.01
2007	1	1	262	0.01	0.02	0.00
2008	1	1	272	0.05	0.07	0.01
2009	1	1	263	0.03	0.04	0.00
2010	1	1	251	0.03	0.05	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 367: Average R-square for Base and Revised Models and FYGPA by Admission Year (Humanidades)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	5	8	236	0.17	0.17	0.00
2005	7	10	412	0.11	0.12	0.01
2006	9	13	514	0.15	0.18	0.03
2007	8	12	534	0.12	0.13	0.00
2008	10	14	513	0.10	0.10	0.00
2009	9	13	506	0.09	0.10	0.00
2010	8	12	385	0.11	0.12	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 368: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ingeniería_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	21	79	3559	0.24	0.25	0.01
2005	21	91	4866	0.23	0.24	0.00
2006	22	102	5596	0.24	0.24	0.00
2007	22	111	6372	0.28	0.29	0.00
2008	22	114	6739	0.26	0.26	0.00
2009	22	118	7467	0.25	0.25	0.00
2010	19	96	6344	0.26	0.26	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 369: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ingeniería_2)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	28	681	0.11	0.12	0.00
2005	13	44	1330	0.12	0.12	0.00
2006	15	56	1719	0.13	0.13	0.00
2007	14	54	1875	0.11	0.12	0.00
2008	16	54	1873	0.13	0.14	0.00
2009	15	58	2018	0.19	0.19	0.00
2010	12	55	2002	0.12	0.12	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 370: Average R-square for Base and Revised Models and FYGPA by Admission Year (Ingeniería_3)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	5	120	0.23	0.27	0.04
2005	3	5	156	0.46	0.46	0.00
2006	4	8	294	0.30	0.30	0.01
2007	6	12	465	0.20	0.20	0.00
2008	6	12	451	0.27	0.28	0.01
2009	6	12	452	0.15	0.15	0.00
2010	6	11	418	0.13	0.13	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 371: Average R-square for Base and Revised Models and FYGPA by Admission Year (Mar)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	5	6	124	0.15	0.16	0.01
2005	8	9	250	0.10	0.11	0.01
2006	7	9	268	0.08	0.08	0.00
2007	6	9	257	0.13	0.13	0.00
2008	8	12	355	0.14	0.14	0.00
2009	7	10	247	0.13	0.13	0.00
2010	7	9	231	0.14	0.15	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 372: Average R-square for Base and Revised Models and FYGPA by Admission Year (Periodismo)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	12	12	450	0.15	0.15	0.00
2005	12	12	590	0.21	0.21	0.00
2006	15	16	750	0.14	0.14	0.00
2007	14	14	757	0.18	0.18	0.00
2008	13	14	654	0.13	0.13	0.00
2009	13	14	667	0.11	0.12	0.00
2010	12	12	567	0.19	0.19	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 373: Average R-square for Base and Revised Models and FYGPA by Admission Year (Salud_1)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	12	20	878	0.39	0.39	0.00
2005	14	26	1466	0.33	0.33	0.01
2006	14	26	1776	0.40	0.41	0.00
2007	14	26	1832	0.35	0.35	0.00
2008	15	28	1849	0.33	0.33	0.00
2009	15	30	2108	0.23	0.23	0.00
2010	14	28	1891	0.30	0.30	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 374: Average R-square for Base and Revised Models and FYGPA by Admission Year (Salud_2)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	20	43	1327	0.19	0.20	0.01
2005	19	45	2140	0.25	0.27	0.02
2006	20	47	2573	0.25	0.25	0.00
2007	19	48	2676	0.27	0.28	0.00
2008	19	48	2780	0.25	0.26	0.01
2009	19	51	2977	0.26	0.26	0.01
2010	19	53	3083	0.24	0.24	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 375: Average R-square for Base and Revised Models and FYGPA by Admission Year (Salud_3)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	17	501	0.28	0.30	0.02
2005	13	21	966	0.40	0.40	0.01
2006	14	23	1172	0.30	0.30	0.00
2007	14	23	1254	0.32	0.32	0.00
2008	14	23	1308	0.23	0.24	0.01
2009	14	23	1276	0.27	0.28	0.00
2010	14	25	1227	0.23	0.23	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 376: Average R-square for Base and Revised Models and FYGPA by Admission Year (Técnico_Administración)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	3	112	0.16	0.21	0.06
2005	5	6	233	0.02	0.03	0.00
2006	5	6	292	0.06	0.08	0.03
2007	5	7	285	0.10	0.10	0.00
2008	5	7	310	0.09	0.11	0.02
2009	5	6	271	0.10	0.10	0.00
2010	6	8	344	0.06	0.06	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 377: Average R-square for Base and Revised Models and FYGPA by Admission Year (Técnico_Agro)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	15	0.22	0.26	0.04
2005	-	-	-	-	-	-
2006	-	-	-	-	-	-
2007	-	-	-	-	-	-
2008	-	-	-	-	-	-
2009	-	-	-	-	-	-
2010	-	-	-	-	-	-

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 378: Average R-square for Base and Revised Models and FYGPA by Admission Year (Técnico_Ciencias)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	27	0.04	0.06	0.02
2005	2	3	76	0.14	0.14	0.01
2006	2	3	90	0.05	0.05	0.00
2007	2	3	77	0.02	0.02	0.00
2008	3	4	123	0.02	0.02	0.00
2009	3	4	114	0.00	0.01	0.01
2010	2	2	49	0.01	0.21	0.20

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 379: Average R-square for Base and Revised Models and FYGPA by Admission Year (Técnico_Diseño)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	16	0.13	0.23	0.10
2005	2	2	45	0.07	0.08	0.00
2006	3	3	70	0.01	0.06	0.04
2007	3	3	69	0.21	0.23	0.01
2008	3	3	68	0.27	0.27	0.00
2009	3	3	69	0.29	0.29	0.00
2010	1	1	28	0.01	0.13	0.12

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 380: Average R-square for Base and Revised Models and FYGPA by Admission Year (Técnico_Educación)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	-	-	-	-	-	-
2005	1	1	22	0.06	0.14	0.09
2006	1	1	24	0.27	0.27	0.00
2007	1	1	18	0.06	0.07	0.01
2008	1	1	20	0.20	0.23	0.03
2009	1	1	27	0.14	0.15	0.01
2010	1	1	25	0.10	0.10	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 381: Average R-square for Base and Revised Models and FYGPA by Admission Year (Técnico_Idioma)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	2	2	58	0.10	0.21	0.11
2005	3	4	133	0.00	0.01	0.01
2006	3	4	148	0.03	0.03	0.00
2007	3	4	156	0.07	0.07	0.00
2008	3	3	138	0.07	0.09	0.02
2009	3	3	128	0.16	0.18	0.02
2010	3	3	99	0.09	0.11	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 382: Average R-square for Base and Revised Models and FYGPA by Admission Year (Técnico_Ingeniería)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	7	20	473	0.05	0.07	0.01
2005	7	40	1176	0.00	0.01	0.01
2006	6	46	1408	0.00	0.01	0.00
2007	7	48	1569	0.02	0.02	0.00
2008	6	48	1840	0.01	0.01	0.00
2009	6	49	1816	0.01	0.01	0.00
2010	4	22	888	0.01	0.01	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 383: Average R-square for Base and Revised Models and FYGPA by Admission Year (Veterinaria)

	Sample Sizes			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	5	262	0.26	0.26	0.00
2005	4	5	393	0.34	0.34	0.01
2006	4	5	407	0.33	0.34	0.01
2007	4	5	454	0.31	0.31	0.01
2008	4	5	440	0.28	0.28	0.00
2009	4	5	419	0.23	0.24	0.00
2010	4	5	408	0.31	0.32	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Appendix T. Incremental Prediction Validity of Ranking by the Type of Career – Second Year Grade Point Average (SYGPA)

Table 384: Average R-square for Base and Revised Models and SYGPA by Admission Year (Administración)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	18	21	1047	0.14	0.15	0.01
2005	19	28	1651	0.14	0.14	0.00
2006	20	28	1856	0.18	0.18	0.00
2007	20	30	2256	0.14	0.14	0.00
2008	20	28	2460	0.09	0.09	0.00
2009	20	29	2578	0.05	0.05	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 385: Average R-square for Base and Revised Models and SYGPA by Admission Year (Administración_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	12	338	0.09	0.12	0.04
2005	14	16	583	0.16	0.19	0.03
2006	14	16	808	0.09	0.13	0.04
2007	15	19	1026	0.12	0.15	0.03
2008	14	17	888	0.10	0.12	0.02
2009	15	18	980	0.14	0.15	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 386: Average R-square for Base and Revised Models and SYGPA by Admission Year (Administración_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	4	86	0.19	0.20	0.01
2005	4	5	137	0.04	0.05	0.00
2006	4	5	253	0.15	0.15	0.00
2007	4	7	252	0.30	0.30	0.00
2008	6	9	335	0.19	0.19	0.00
2009	6	10	338	0.15	0.15	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 387: Average R-square for Base and Revised Models and SYGPA by Admission Year (Agro)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	8	10	462	0.21	0.22	0.01
2005	8	12	717	0.19	0.20	0.00
2006	8	12	824	0.25	0.25	0.01
2007	8	14	860	0.21	0.22	0.01
2008	8	12	797	0.21	0.22	0.01
2009	8	14	814	0.24	0.24	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 388: Average R-square for Base and Revised Models and SYGPA by Admission Year (Agro_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	2	2	46	0.03	0.08	0.05
2005	3	4	90	0.02	0.04	0.02
2006	3	4	75	0.05	0.10	0.04
2007	3	3	81	0.15	0.19	0.04
2008	1	1	41	0.01	0.05	0.04
2009	2	2	58	0.13	0.15	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 389: Average R-square for Base and Revised Models and SYGPA by Admission Year (Arquitectura)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	10	505	0.18	0.18	0.00
2005	12	12	708	0.15	0.15	0.00
2006	12	12	778	0.09	0.09	0.00
2007	12	12	774	0.16	0.16	0.00
2008	12	13	750	0.15	0.15	0.00
2009	13	14	780	0.23	0.23	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 390: Average R-square for Base and Revised Models and SYGPA by Admission Year (Arte_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	4	160	0.10	0.10	0.00
2005	5	6	245	0.05	0.06	0.00
2006	8	6	336	0.08	0.08	0.00
2007	6	8	312	0.12	0.13	0.01
2008	6	7	301	0.16	0.18	0.01
2009	7	8	312	0.14	0.14	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 391: Average R-square for Base and Revised Models and SYGPA by Admission Year (Arte_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	2	61	0.29	0.31	0.03
2005	5	9	212	0.12	0.12	0.00
2006	4	9	174	0.11	0.11	0.00
2007	5	8	200	0.12	0.13	0.01
2008	4	6	184	0.07	0.07	0.00
2009	5	6	190	0.16	0.16	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 392: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ciencias)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	6	8	269	0.12	0.12	0.00
2005	6	10	366	0.13	0.14	0.01
2006	7	10	437	0.12	0.12	0.00
2007	8	15	541	0.06	0.06	0.00
2008	8	15	521	0.02	0.02	0.00
2009	6	11	467	0.06	0.07	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 393: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ciencias_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	18	541	0.15	0.15	0.00
2005	11	23	806	0.12	0.12	0.00
2006	11	23	902	0.15	0.15	0.00
2007	12	24	964	0.12	0.12	0.00
2008	12	24	952	0.12	0.13	0.01
2009	12	24	863	0.08	0.08	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 394: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ciencias_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	5	134	0.18	0.21	0.03
2005	4	7	211	0.18	0.18	0.00
2006	6	7	259	0.24	0.25	0.00
2007	6	8	223	0.34	0.34	0.00
2008	4	5	132	0.30	0.30	0.00
2009	4	5	112	0.17	0.18	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 395: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ciencias_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	-	-	-	-	-	-
2005	1	1	27	0.03	0.21	0.18
2006	2	1	46	0.20	0.21	0.01
2007	2	4	102	0.13	0.13	0.01
2008	2	4	133	0.30	0.30	0.00
2009	2	4	138	0.35	0.35	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 396: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ciencias_Sociales_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	14	21	679	0.21	0.21	0.00
2005	17	29	1193	0.15	0.16	0.00
2006	17	29	1337	0.26	0.26	0.00
2007	17	28	1374	0.20	0.20	0.01
2008	16	27	1397	0.23	0.23	0.00
2009	16	28	1385	0.22	0.22	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 397: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ciencias_Sociales_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	6	11	369	0.19	0.20	0.01
2005	9	17	680	0.21	0.22	0.00
2006	12	17	848	0.19	0.19	0.00
2007	13	27	1109	0.18	0.19	0.00
2008	13	27	1159	0.23	0.23	0.00
2009	13	28	1294	0.15	0.15	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 398: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ciencias_Sociales_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	11	245	0.09	0.12	0.04
2005	11	11	388	0.13	0.15	0.03
2006	12	11	544	0.10	0.10	0.01
2007	12	14	566	0.07	0.07	0.00
2008	11	13	520	0.06	0.06	0.00
2009	12	14	590	0.04	0.05	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 399: Average R-square for Base and Revised Models and SYGPA by Admission Year (Comunicaciones)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	22	0.24	0.26	0.02
2005	1	1	33	0.10	0.24	0.14
2006	2	1	59	0.01	0.11	0.10
2007	2	2	69	0.01	0.01	0.00
2008	2	2	61	0.05	0.05	0.00
2009	1	1	38	0.07	0.09	0.03

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 400: Average R-square for Base and Revised Models and SYGPA by Admission Year (Construcción)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	9	9	359	0.15	0.15	0.00
2005	11	11	448	0.06	0.06	0.00
2006	10	11	535	0.10	0.10	0.00
2007	12	12	649	0.12	0.13	0.01
2008	12	12	643	0.07	0.07	0.01
2009	13	13	691	0.08	0.09	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 401: Average R-square for Base and Revised Models and SYGPA by Admission Year (Derecho)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	15	17	1032	0.15	0.15	0.00
2005	14	18	1375	0.16	0.16	0.00
2006	15	18	1613	0.18	0.18	0.00
2007	15	19	1713	0.15	0.15	0.00
2008	14	18	1808	0.17	0.17	0.00
2009	14	19	1697	0.23	0.24	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 402: Average R-square for Base and Revised Models and SYGPA by Admission Year (Diseño)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	8	10	388	0.09	0.09	0.00
2005	11	15	635	0.06	0.07	0.00
2006	12	15	850	0.08	0.08	0.00
2007	11	17	847	0.05	0.05	0.00
2008	11	15	825	0.07	0.07	0.00
2009	11	16	731	0.09	0.09	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 403: Average R-square for Base and Revised Models and SYGPA by Admission Year (Educación)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	13	25	496	0.08	0.10	0.02
2005	16	49	1287	0.10	0.10	0.00
2006	17	49	1582	0.09	0.09	0.00
2007	17	55	1767	0.08	0.08	0.00
2008	17	55	1737	0.09	0.09	0.00
2009	17	56	1725	0.09	0.09	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 404: Average R-square for Base and Revised Models and SYGPA by Admission Year (Educación_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	12	24	572	0.17	0.17	0.00
2005	14	42	1364	0.08	0.09	0.01
2006	14	42	1718	0.11	0.11	0.00
2007	16	51	1914	0.09	0.09	0.00
2008	15	46	1862	0.05	0.05	0.00
2009	15	45	1947	0.07	0.07	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 405: Average R-square for Base and Revised Models and SYGPA by Admission Year (Educación_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	9	11	232	0.08	0.08	0.01
2005	11	27	765	0.09	0.09	0.00
2006	12	27	956	0.11	0.11	0.00
2007	11	28	888	0.05	0.05	0.00
2008	12	27	843	0.08	0.08	0.00
2009	12	31	910	0.08	0.08	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 406: Average R-square for Base and Revised Models and SYGPA by Admission Year (Educación_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	8	10	228	0.10	0.12	0.02
2005	8	13	445	0.18	0.19	0.00
2006	11	13	633	0.13	0.13	0.00
2007	11	17	638	0.06	0.06	0.00
2008	11	17	658	0.07	0.07	0.00
2009	11	17	690	0.08	0.08	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 407: Average R-square for Base and Revised Models and SYGPA by Admission Year (General)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	181	0.08	0.08	0.00
2005	1	1	220	0.05	0.06	0.01
2006	1	1	237	0.08	0.08	0.00
2007	1	1	248	0.05	0.05	0.00
2008	1	1	247	0.04	0.04	0.00
2009	1	1	192	0.02	0.02	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 408: Average R-square for Base and Revised Models and SYGPA by Admission Year (Humanidades)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	7	207	0.10	0.11	0.00
2005	7	10	372	0.12	0.12	0.00
2006	6	10	406	0.15	0.17	0.02
2007	7	10	449	0.16	0.16	0.00
2008	8	12	429	0.08	0.08	0.00
2009	9	13	448	0.10	0.10	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 409: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ingeniería_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	21	77	3199	0.22	0.23	0.00
2005	21	89	4384	0.20	0.20	0.00
2006	21	89	4783	0.23	0.23	0.00
2007	21	106	5414	0.22	0.22	0.00
2008	21	107	5558	0.20	0.20	0.00
2009	22	100	5800	0.21	0.21	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 410: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ingeniería_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	9	22	485	0.11	0.11	0.00
2005	13	38	1018	0.10	0.10	0.00
2006	15	38	1300	0.11	0.12	0.00
2007	13	46	1400	0.07	0.07	0.00
2008	15	49	1441	0.08	0.09	0.00
2009	15	53	1506	0.10	0.10	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 411: Average R-square for Base and Revised Models and SYGPA by Admission Year (Ingeniería_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	5	115	0.20	0.23	0.04
2005	4	6	171	0.23	0.24	0.01
2006	4	6	251	0.17	0.18	0.01
2007	6	11	406	0.20	0.20	0.00
2008	6	10	370	0.16	0.18	0.02
2009	6	11	353	0.09	0.10	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 412: Average R-square for Base and Revised Models and SYGPA by Admission Year (Mar)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	5	6	113	0.18	0.18	0.00
2005	7	9	204	0.16	0.16	0.00
2006	7	9	191	0.10	0.11	0.01
2007	4	5	152	0.23	0.24	0.00
2008	8	9	226	0.15	0.15	0.00
2009	5	7	142	0.16	0.19	0.03

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 413: Average R-square for Base and Revised Models and SYGPA by Admission Year (Periodismo)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	10	398	0.22	0.23	0.00
2005	12	12	537	0.15	0.15	0.00
2006	13	12	647	0.13	0.14	0.00
2007	14	14	683	0.22	0.22	0.00
2008	12	12	545	0.17	0.17	0.00
2009	12	12	548	0.19	0.19	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 414: Average R-square for Base and Revised Models and SYGPA by Admission Year (Salud_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	13	20	863	0.27	0.28	0.00
2005	15	27	1467	0.36	0.37	0.00
2006	14	27	1713	0.27	0.27	0.00
2007	14	26	1767	0.14	0.14	0.00
2008	15	26	1649	0.03	0.03	0.00
2009	14	27	1797	0.08	0.08	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 415: Average R-square for Base and Revised Models and SYGPA by Admission Year (Salud_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	20	41	1241	0.16	0.17	0.01
2005	21	49	2131	0.19	0.20	0.01
2006	20	49	2451	0.20	0.20	0.00
2007	20	50	2544	0.14	0.15	0.01
2008	20	49	2554	0.11	0.11	0.00
2009	20	50	2480	0.12	0.12	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 416: Average R-square for Base and Revised Models and SYGPA by Admission Year (Salud_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	16	469	0.14	0.14	0.00
2005	14	22	942	0.18	0.19	0.01
2006	14	22	1064	0.13	0.14	0.00
2007	15	24	1158	0.08	0.08	0.00
2008	15	24	1143	0.08	0.08	0.00
2009	15	22	980	0.13	0.13	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 417: Average R-square for Base and Revised Models and SYGPA by Admission Year (Técnico_Administración)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	2	2	87	0.14	0.25	0.11
2005	5	6	195	0.16	0.16	0.00
2006	5	6	253	0.04	0.05	0.01
2007	5	7	246	0.06	0.06	0.00
2008	5	7	276	0.06	0.07	0.02
2009	5	6	227	0.04	0.05	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 418: Average R-square for Base and Revised Models and SYGPA by Admission Year (Técnico_Ciencias)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	24	0.07	0.10	0.03
2005	2	3	74	0.12	0.22	0.10
2006	2	3	86	0.04	0.07	0.03
2007	2	3	70	0.01	0.01	0.00
2008	3	4	116	0.05	0.08	0.03
2009	3	4	109	0.07	0.07	0.01

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 419: Average R-square for Base and Revised Models and SYGPA by Admission Year (Técnico_Diseño)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	15	0.07	0.09	0.01
2005	2	2	41	0.02	0.03	0.00
2006	3	2	61	0.01	0.03	0.02
2007	3	3	62	0.09	0.09	0.00
2008	2	2	49	0.17	0.17	0.00
2009	2	2	48	0.02	0.02	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 420: Average R-square for Base and Revised Models and SYGPA by Admission Year (Técnico_Educación)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	-	-	-	-	-	-
2005	1	1	19	0.07	0.11	0.03
2006	1	1	18	0.22	0.25	0.04
2007	-	-	-	-	-	-
2008	1	1	15	0.18	0.25	0.06
2009	1	1	21	0.06	0.08	0.02

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 421: Average R-square for Base and Revised Models and SYGPA by Admission Year (Técnico_Idioma)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	2	2	57	0.17	0.17	0.00
2005	3	3	110	0.15	0.15	0.00
2006	3	3	120	0.18	0.25	0.07
2007	3	3	125	0.09	0.10	0.01
2008	3	3	129	0.07	0.08	0.01
2009	3	3	114	0.02	0.05	0.03

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 422: Average R-square for Base and Revised Models and SYGPA by Admission Year (Técnico_Ingeniería)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	15	358	0.07	0.07	0.00
2005	4	37	1031	0.02	0.03	0.01
2006	6	37	1239	0.02	0.02	0.00
2007	5	46	1356	0.03	0.03	0.00
2008	6	46	1629	0.02	0.03	0.00
2009	6	48	1585	0.01	0.01	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Table 423: Average R-square for Base and Revised Models and SYGPA by Admission Year (Veterinaria)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	5	239	0.10	0.11	0.01
2005	4	5	359	0.16	0.17	0.01
2006	4	5	377	0.09	0.09	0.00
2007	4	5	413	0.13	0.15	0.01
2008	4	5	378	0.18	0.18	0.00
2009	4	5	355	0.21	0.21	0.00

Note: Base Model: Predictor variables= PSU tests and NEM
Revised Model: Predictor variables= Base Model and Ranking
Difference: Revised Model minus Base Model

Appendix U. Incremental Prediction Validity of Ranking by the Type of Career - University Completion

Table 424: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Administración)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	19	27	1345	0.09	0.09	0.00
2005	20	30	2055	0.08	0.08	0.00
2006	21	31	2300	0.01	0.02	0.01

Table 425: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Administración_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	14	15	508	0.01	0.01	0.00
2005	16	18	782	0.04	0.04	0.01
2006	16	21	1057	0.02	0.03	0.01

Table 426: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Administración_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	4	103	0.00	0.02	0.02
2005	4	6	190	0.01	0.01	0.00
2006	4	7	300	0.03	0.03	0.00

Table 427: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Agro)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	8	12	560	0.07	0.07	0.01
2005	8	13	854	0.04	0.04	0.00
2006	8	14	982	0.03	0.03	0.00

Table 428: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Agro_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	2	2	54	0.09	0.12	0.03
2005	3	4	110	0.01	0.02	0.01
2006	3	3	91	0.00	0.01	0.01

Table 429: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Arquitectura)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	17	17	875	0.06	0.06	0.00
2005	18	18	1168	0.01	0.01	0.00
2006	15	15	1146	0.00	0.00	0.00

Table 430: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Arte_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	4	178	0.03	0.03	0.00
2005	7	9	352	0.12	0.12	0.00
2006	8	9	398	0.11	0.11	0.00

Table 431: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Arte_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	4	100	0.10	0.17	0.07
2005	5	9	254	0.03	0.04	0.01
2006	5	7	234	0.04	0.05	0.01

Table 432: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ciencias)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	7	9	356	0.06	0.06	0.00
2005	7	12	514	0.05	0.05	0.00
2006	9	15	624	0.05	0.05	0.00

Table 433: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ciencias_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	23	695	0.09	0.09	0.00
2005	11	23	950	0.09	0.09	0.00
2006	12	24	1077	0.05	0.05	0.00

Table 434: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ciencias_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	5	8	239	0.03	0.04	0.01
2005	7	11	369	0.00	0.01	0.01
2006	7	12	409	0.04	0.04	0.00

Table 435: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ciencias_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	47	0.05	0.05	0.00
2005	2	2	85	0.02	0.06	0.04
2006	2	2	92	-	-	-

Table 436: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ciencias_Sociales_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	17	24	899	0.14	0.14	0.00
2005	19	31	1431	0.15	0.15	0.00
2006	20	32	1666	0.14	0.14	0.00

Table 437: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ciencias_Sociales_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	6	13	525	0.12	0.14	0.02
2005	10	18	854	0.06	0.08	0.02
2006	12	20	1074	0.09	0.10	0.01

Table 438: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ciencias_Sociales_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	13	13	348	0.02	0.04	0.02
2005	13	13	532	0.01	0.03	0.02
2006	12	14	689	0.05	0.06	0.01

Table 439: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Comunicaciones)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	32	0.00	0.02	0.02
2005	1	1	41	0.00	0.01	0.01
2006	2	2	76	0.09	0.13	0.04

Table 440: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Construcción)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	10	624	0.13	0.15	0.02
2005	14	14	874	0.08	0.08	0.00
2006	15	15	1051	0.01	0.01	0.00

Table 441: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Derecho)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	17	20	1432	0.19	0.19	0.00
2005	17	21	1964	0.17	0.17	0.00
2006	17	21	2176	0.08	0.08	0.00

Table 442: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Diseño)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	13	542	0.03	0.03	0.00
2005	13	18	836	0.08	0.09	0.01
2006	13	19	1043	0.03	0.04	0.01

Table 443: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Educación)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	16	32	711	0.03	0.04	0.01
2005	18	56	1712	0.04	0.05	0.01
2006	19	60	2062	0.01	0.01	0.00

Table 444: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Educación_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	13	28	679	0.04	0.04	0.00
2005	15	46	1561	0.02	0.02	0.00
2006	16	52	1990	0.04	0.04	0.00

Table 445: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Educación_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	10	18	654	0.09	0.09	0.00
2005	12	32	1389	0.04	0.04	0.00
2006	13	35	1634	0.02	0.02	0.00

Table 446: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Educación_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	9	11	273	0.01	0.01	0.00
2005	8	13	485	0.07	0.07	0.00
2006	11	17	714	0.02	0.03	0.01

Table 447: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (General)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	194	0.04	0.04	0.00
2005	1	1	247	0.01	0.02	0.01
2006	1	1	257	0.00	0.00	0.00

Table 448: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Humanidades)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	5	8	245	0.09	0.09	0.00
2005	7	10	432	0.07	0.07	0.00
2006	9	13	558	0.05	0.05	0.00

Table 449: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ingeniería_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	21	84	3997	0.08	0.08	0.00
2005	22	97	5617	0.06	0.06	0.00
2006	23	108	6319	0.03	0.03	0.00

Table 450: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ingeniería_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	15	36	944	0.00	0.00	0.00
2005	15	50	1666	0.00	0.00	0.00
2006	16	61	2072	0.00	0.01	0.01

Table 451: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Ingeniería_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	3	5	129	0.34	0.35	0.01
2005	4	6	196	0.18	0.18	0.00
2006	4	8	300	0.08	0.08	0.00

Table 452: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Mar)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	8	9	189	0.01	0.01	0.00
2005	9	14	407	0.06	0.06	0.00
2006	9	11	361	0.03	0.03	0.00

Table 453: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Periodismo)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	13	13	496	0.07	0.09	0.02
2005	13	13	648	0.11	0.11	0.00
2006	15	16	817	0.13	0.14	0.01

Table 454: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Salud_1)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	14	23	978	0.14	0.14	0.00
2005	15	27	1546	0.02	0.02	0.00
2006	15	27	1859	0.04	0.05	0.01

Table 455: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Salud_2)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	20	47	1473	0.07	0.07	0.00
2005	21	53	2505	0.08	0.09	0.01
2006	21	55	2987	0.08	0.08	0.00

Table 456: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Salud_3)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	11	19	581	0.07	0.07	0.00
2005	14	22	1072	0.12	0.12	0.00
2006	15	24	1248	0.13	0.13	0.00

Table 457: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Técnico_Administración)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	4	142	0.01	0.02	0.01
2005	5	6	255	0.01	0.01	0.00
2006	5	6	312	0.02	0.02	0.00

Table 458: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Técnico_Agro)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	20	0.23	0.25	0.02
2005	N/A	N/A	N/A	N/A	N/A	N/A
2006	N/A	N/A	N/A	N/A	N/A	N/A

Table 459: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Técnico_Ciencias)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	27	0.03	0.05	0.02
2005	2	3	78	0.04	0.04	0.00
2006	2	3	94	0.05	0.08	0.03

Table 460: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Técnico_Diseño)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	1	1	17	0.30	0.31	0.01
2005	2	2	51	0.18	0.20	0.02
2006	3	3	73	0.04	0.07	0.03

Table 461: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Técnico_Educación)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	N/A	N/A	N/A	N/A	N/A	N/A
2005	1	1	22	0.27	0.32	0.05
2006	1	1	27	0.03	0.04	0.01

Table 462: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Técnico_Idioma)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	4	172	0.01	0.02	0.01
2005	5	7	296	0.01	0.02	0.01
2006	4	6	311	0.02	0.02	0.00

Table 463: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Técnico_Ingeniería)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	9	25	585	0.05	0.06	0.01
2005	7	41	1275	0.00	0.01	0.01
2006	6	47	1514	0.01	0.01	0.00

Table 464: Average R-square (Cox-Snell) for Base and Revised Models and University Completion by Admission Year (Veterinaria)

	Sample Size			Models		Difference
Year	University	Career	Student	Base	Revised	
2004	4	5	266	0.17	0.17	0.00
2005	4	5	408	0.24	0.24	0.00
2006	4	5	423	0.07	0.07	0.00

Appendix V. Prediction Bias by the Type of Career – First Year Grade Point Average (FYGPA)

Table 465: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Administración)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	35	9430	-0.12	-0.1	-0.12	-0.13	-0.03	-0.07
Female	35	7464	0.15	0.13	0.15	0.16	0.04	0.07
SES:								
QA	35	1961	0.08	0.06	0.07	0.08	0	0
QB	35	1452	-	-0.02	-0.04	-0.02	-0.06	-0.13
QC	35	1303	-0.02	-0.06	-0.04	-0.06	-0.05	-0.1
QD	35	1132	0.02	0.01	0.04	-0.02	0.07	-0.01
QE	35	2256	-0.13	-0.11	-0.13	0	-0.06	-0.05
Curricular Branch:								
Scientific-Humanistic	35	14993	-0.06	-0.04	-0.04	-0.03	-0.02	-
Technical-Professional	35	1900	0.23	0.15	0.12	0.12	0.07	-0.03
High School Type:								
Municipal	35	3586	0.05	0.03	0.04	0.03	0.02	-0.06
Subsidized	35	5295	-0.02	-0.02	-0.04	-	-0.03	-0.03
Private	34	8012	-0.11	-0.06	-0.03	-0.06	0.07	0.11
Region:								
Center	34	10495	-0.04	-0.05	-0.06	-0.01	-0.01	-0.07
North	27	1111	-0.1	-0.06	-0.08	-0.04	-0.15	-0.07
South	35	5287	0.01	0	0.01	0.05	0.01	-0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 466: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Administración_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	26	2823	-0.14	-0.09	-0.14	-0.12	-0.01	-0.09
Female	26	3381	0.13	0.1	0.14	0.11	0.03	0.05
SES:								
QA	26	1087	0.15	0.12	0.12	0.11	0.05	-
QB	26	1327	0.04	0.04	0.01	0.1	0.01	-0.01
QC	25	888	-0.15	-0.13	-0.11	-0.07	-0.12	-0.15
QD	22	548	0.04	0.08	0.07	-0.07	0.13	0.13
QE	22	511	-0.36	-0.3	-0.32	-0.1	-0.21	-0.03
Curricular Branch:								
Scientific-Humanistic	26	3531	-0.23	-0.17	-0.16	-0.04	-0.13	-0.1
Technical-Professional	26	2670	0.21	0.17	0.16	0.09	0.12	0.04
High School Type:								
Municipal	26	2998	0.01	0.02	0.02	0.02	0.01	0
Subsidized	26	2750	-0.01	-0.02	-0.03	0.01	-	-0.01
Private	18	453	-0.31	-0.21	-0.2	-0.01	0.02	0.28
Region:								
Center	24	3200	-0.05	-0.11	-0.18	-0.06	-0.01	-0.1
North	12	197	-0.27	-0.2	-0.4	0.2	-0.19	-0.33
South	26	2804	-0.03	-0.01	0.01	-	-0.02	-0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 467: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Administración_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	10	802	-0.17	-0.14	-0.16	-0.02	-0.06	0.05
Female	10	1326	0.12	0.09	0.11	0.08	-0.02	-0.02
SES:								
QA	10	226	-0.01	0.01	-0.05	-	-0.02	0.08
QB	10	258	0.03	-0.02	-0.2	0.32	-0.07	-0.12
QC	10	269	0.16	0.05	0	0.11	0.03	-0.28
QD	10	215	0.08	0.08	0.12	0.04	0.07	0.08
QE	10	312	-0.11	-0.07	-0.11	0.04	-0.02	-
Curricular Branch:								
Scientific-Humanistic	10	1801	-0.03	-0.01	-0.02	-0.02	-0.03	0.04
Technical-Professional	10	325	0.03	-0.04	-0.09	0.21	-0.15	-0.33
High School Type:								
Municipal	10	510	0.06	-0.01	-0.06	0.11	-0.02	-0.05
Subsidized	10	1108	0.05	0.05	0.06	0.01	-0.06	-0.09
Private	10	508	-0.11	-0.06	-0.01	0.01	0.01	0.13
Region:								
Center	9	1541	-0.02	-0.03	-0.03	-0.04	-0.04	-0.03
North	10	176	-0.1	-0.11	-0.04	-0.04	-0.16	0.09
South	9	409	0.1	0.1	0.01	0.03	0.1	0.25

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 468: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Agro)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	16	3480	-0.07	-0.05	-	-0.07	-0.01	-0.04
Female	16	2598	0.1	0.06	0.02	0.1	0.01	0.06
SES:								
QA	16	738	0.07	0.05	0.09	0.06	-0.01	0.02
QB	16	805	-0.05	-0.07	-0.06	-0.06	-0.08	-0.13
QC	16	714	0.03	0.02	-0.05	0.03	0.01	-0.01
QD	16	584	0.03	0.01	0.13	0.01	0.02	0.01
QE	16	743	-0.06	0.01	0.01	-0.01	0.01	0.07
Curricular Branch:								
Scientific-Humanistic	16	5481	-0.03	-0.02	0	-0.02	-0.02	0
Technical-Professional	16	596	0.04	-0.03	-0.06	0.02	-0.04	-0.18
High School Type:								
Municipal	16	2066	0.03	-	0.06	0.02	-0.01	-0.06
Subsidized	16	2470	-0.01	-0.01	-0.08	-0.02	-0.02	-0.01
Private	15	1541	-0.02	0.09	0.17	0.01	0.1	0.16
Region:								
Center	16	3089	0.11	0.11	0.08	0.08	0.17	0.05
North	14	156	0.09	0.09	-0.34	0.1	0.16	-0.1
South	16	2832	-0.1	-0.09	-0.12	-0.08	-0.1	-0.08

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 469: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Agro_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	4	301	-0.15	-0.11	-0.2	-0.13	-0.08	-0.07
Female	4	351	0.17	0.12	0.2	0.13	0.08	0.07
SES:								
QA	4	87	0.13	0.11	0.28	0.21	0.09	0.03
QB	4	144	0.03	0.02	0.05	0.03	-0.02	-0.01
QC	4	87	-0.2	-0.19	0	-0.18	-0.13	-0.15
QD	4	74	0.03	0.08	-0.4	0.05	-0.03	-0.07
QE	4	52	-0.41	-0.35	-0.5	-0.38	-0.15	-0.15
Curricular Branch:								
Scientific-Humanistic	4	532	-0.05	-0.01	0.02	-0.02	-	0.01
Technical-Professional	4	120	0.13	0.02	-0.05	0.06	-0.05	-0.09
High School Type:								
Municipal	4	235	-0.05	-0.09	-0.04	-0.04	-0.1	-0.12
Subsidized	4	368	0.07	0.09	0.05	0.06	0.07	0.08
Private	4	49	0.14	0.21	-0.02	0.13	0.33	0.42
Region:								
Center	4	408	-0.01	0	0.09	-0.13	0.08	-0.07
North	3	7	0.16	0.12	0.11	0.2	0.08	-
South	4	237	0.13	0.13	0.09	0.22	0.11	0.11

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 470: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Arquitectura)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	19	2977	-0.13	-0.11	-0.15	-0.09	-0.09	-0.08
Female	19	2923	0.18	0.13	0.17	0.12	0.09	0.09
SES:								
QA	18	539	0.03	0.05	0.05	0.01	0.05	-0.1
QB	19	391	0.04	-0.05	-	-	-0.05	-0.18
QC	19	448	-0.03	-0.08	-0.04	-0.1	-0.05	-0.09
QD	19	452	-0.05	-0.08	-0.05	-0.09	-0.09	-0.11
QE	16	775	-0.03	0.05	0.05	-0.02	0.03	0.07
Curricular Branch:								
Scientific-Humanistic	19	5513	0.01	0.01	0.02	0.01	0.01	0.02
Technical-Professional	18	386	-0.03	-0.11	-0.1	-0.16	-0.08	-0.25
High School Type:								
Municipal	19	1171	-0.03	-0.1	-0.08	0	-0.07	-0.21
Subsidized	19	2085	0.02	0.02	0.03	0.06	-0.01	-0.03
Private	19	2643	0.04	0.07	0.07	0.04	0.16	0.21
Region:								
Center	18	3213	0.04	0.01	-0.01	0.11	0.09	-0.01
North	15	542	-0.06	-0.08	-0.08	0.01	-0.19	-0.07
South	19	2144	-0.13	-0.12	-0.08	-0.05	-0.11	-0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 471: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Arte_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	14	549	-0.24	-0.21	-0.22	-0.33	-0.09	-0.12
Female	14	1952	0.09	0.08	0.09	0.22	0.06	0.05
SES:								
QA	13	228	-0.04	0.01	-	0.16	-	0.06
QB	14	161	-0.21	-0.27	-0.3	-0.08	-0.17	-0.28
QC	14	173	-0.26	-0.25	-0.27	0.08	-0.4	-0.37
QD	14	184	-0.09	-0.14	-0.13	-0.27	-0.2	-0.18
QE	14	347	0.13	0.13	0.16	-0.02	0.09	0.1
Curricular Branch:								
Scientific-Humanistic	14	2420	0.02	0.02	0.02	0.04	0.03	0.01
Technical-Professional	12	81	-0.11	-0.13	-0.16	-0.28	-0.26	-0.38
High School Type:								
Municipal	14	425	-0.13	-0.08	-0.09	0.11	-0.08	-0.15
Subsidized	14	831	-0.02	-0.03	-0.05	-0.01	-0.08	-0.12
Private	14	1245	0.08	0.1	0.13	0.22	0.15	0.2
Region:								
Center	14	1805	-0.05	-0.06	-0.04	-0.34	-	0.01
North	14	87	0.03	0.09	0.1	-0.44	-0.02	0.1
South	14	609	-0.22	-0.23	-0.24	-0.01	-0.2	-0.23

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 472: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Arte_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	11	704	-0.04	-0.02	-0.06	-0.02	0.01	-0.03
Female	11	806	0.07	0.03	0.09	0.01	0.01	0.07
SES:								
QA	11	112	0.12	0.14	0.16	0.42	0.16	0.1
QB	11	121	-0.13	-0.05	-0.16	-0.11	-0.07	-0.23
QC	11	129	-0.27	-0.25	-0.3	-0.38	-0.25	-0.3
QD	11	112	-0.03	-0.06	-0.07	-0.37	0.06	0.01
QE	10	202	0.01	-0.03	-0.01	0.03	-0.03	0.01
Curricular Branch:								
Scientific-Humanistic	11	1435	0.01	0.01	0.01	-0.01	0.01	0.01
Technical-Professional	11	75	-0.21	-0.21	-0.37	0.14	-0.19	-0.05
High School Type:								
Municipal	11	269	-	0.02	-0.01	0.06	0.04	-
Subsidized	11	544	0.02	-0.01	-0.02	-0.18	-0.02	-0.06
Private	11	697	-0.01	-0.04	0.01	-	-0.03	0.15
Region:								
Center	11	1153	0.04	0.01	0.04	-0.04	0.09	0.06
North	9	75	0	-0.02	0.04	-0.46	-0.18	-0.05
South	11	282	0.02	0	-0.03	0.35	-0.05	0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 473: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ciencias)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	20	2285	-0.11	-0.05	-0.05	-0.11	-0.02	-0.02
Female	20	2767	0.09	0.05	0.06	0.1	0.02	0.01
SES:								
QA	19	393	0.05	0.02	-0.2	0.03	0.02	-0.01
QB	19	531	-0.01	-0.06	-0.04	-0.02	-0.11	-0.03
QC	20	548	0.17	0.13	0.12	0.05	0.12	-0.02
QD	20	456	0.06	0.11	0.22	0.02	0.07	0.04
QE	20	679	-0.02	0.03	-0.12	-0.05	0.04	-0.06
Curricular Branch:								
Scientific-Humanistic	20	4691	-	0.01	0	0	0.02	0.02
Technical-Professional	20	360	-0.05	-0.14	-0.01	-0.04	-0.14	-0.28
High School Type:								
Municipal	20	1368	0.04	0.02	-0.13	0.04	0.04	-0.01
Subsidized	20	2294	0.05	0.05	0.14	0.05	0.02	0.01
Private	20	1389	-0.05	0.01	-0.11	-0.04	0.08	0.21
Region:								
Center	18	3763	0.14	0.15	0.14	0.16	0.16	0.18
North	18	278	0.03	0.04	0.14	-0.01	-0.03	-0.08
South	18	1010	0	0	0.06	-0.02	-0.06	-0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 474: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ciencias_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	26	2980	-0.07	-0.04	-	-0.08	0.02	-
Female	26	3738	0.05	0.03	0	0.07	-0.02	-
SES:								
QA	25	601	0.01	0.01	0.41	-0.02	-0.04	-0.02
QB	26	981	-0.02	-0.04	0.1	-0.04	-0.07	-0.11
QC	26	907	0.01	-	0.07	0.01	-0.01	-0.02
QD	26	756	-0.05	-0.05	-0.12	-0.05	-0.03	-0.03
QE	26	797	0.03	0.05	-0.21	0.05	0.07	0.09
Curricular Branch:								
Scientific-Humanistic	26	6433	0	0.01	0	0.01	0.01	0.01
Technical-Professional	26	284	-0.07	-0.13	-0.19	-0.12	-0.14	-0.26
High School Type:								
Municipal	26	2349	-	-0.02	-0.02	-0.02	-0.01	-0.06
Subsidized	26	3382	0.01	0.01	0.02	0.01	-	0
Private	26	986	-0.17	-0.09	0.14	-0.14	-0.01	0.13
Region:								
Center	26	3658	-0.01	-0.02	0.09	-0.01	0.01	0.01
North	26	623	-0.01	0.01	0.02	0.05	-0.05	0
South	25	2436	-0.09	-0.08	0.15	-0.09	-0.09	-0.1

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 475: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ciencias_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	13	1113	-	0.02	0.01	-0.01	0.05	0.04
Female	13	781	0.02	-0.02	-0.01	0.06	-0.06	-0.04
SES:								
QA	13	220	-0.05	-0.1	0.04	-0.07	-0.11	-0.23
QB	12	292	-0.12	-0.13	0.01	-0.11	-0.21	-0.18
QC	13	248	-0.12	-0.1	-0.23	-0.09	-0.08	-0.11
QD	13	169	0.03	0.04	-0.3	0.03	-	0.04
QE	13	193	0.06	0.1	0.13	0.06	0.11	0.21
Curricular Branch:								
Scientific-Humanistic	13	1714	-0.03	-0.02	0.02	-0.01	-0.02	0.02
Technical-Professional	12	179	0.18	0.11	-0.39	0.12	0.17	-0.21
High School Type:								
Municipal	13	669	0.01	-0.02	0.2	-0.01	-0.01	-0.03
Subsidized	13	882	0	-	-0.09	-0.01	-0.03	-0.03
Private	13	342	0.03	0.14	0.06	0.18	0.29	0.45
Region:								
Center	12	1017	-0.03	-0.09	-0.07	-0.04	0.12	0.11
North	13	198	-0.06	-0.03	0.25	-0.03	-0.08	-0.08
South	13	679	0.13	0.18	0.04	0.12	0.11	0

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 476: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ciencias_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	4	426	-	0.06	0.04	0.02	0.07	0.1
Female	4	249	-	-0.11	-0.06	-0.03	-0.13	-0.17
SES:								
QA	4	46	-0.13	-0.12	-0.09	-0.06	-0.11	-0.16
QB	4	57	0.25	0.17	0.22	0.26	0.13	0.12
QC	4	71	-0.02	-0.05	0.16	-0.02	-0.07	-0.06
QD	4	50	-0.17	-0.19	0.18	-0.15	-0.23	-0.1
QE	4	75	0.04	0.04	-0.41	-	0.03	0.02
Curricular Branch:								
Scientific-Humanistic	4	641	-0.02	-0.01	0.02	-0.02	-0.01	-0.01
Technical-Professional	4	34	0.35	0.19	-0.13	0.28	0.16	0.15
High School Type:								
Municipal	4	204	0.05	-0.02	0.17	0.01	-0.07	-0.12
Subsidized	4	338	-0.01	-	-0.02	0.01	0.01	0
Private	4	133	-0.1	0.03	-0.26	-0.05	0.08	0.18
Region:								
Center	4	68	-0.01	-0.04	0.02	-0.11	0.04	-0.08
North	4	237	-0.11	-0.02	-0.05	-0.07	-0.11	-0.11
South	4	370	-0.05	-0.03	0.03	-0.04	-0.02	-0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 477: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ciencias_Sociales_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	32	3611	-0.21	-0.19	-0.21	-0.21	-0.09	-0.14
Female	32	6015	0.14	0.11	0.13	0.12	0.05	0.08
SES:								
QA	32	934	0.04	0.01	0.03	0.04	0.01	0.03
QB	32	850	-0.01	-0.06	-0.06	-0.11	-0.09	-0.15
QC	32	830	-0.05	-0.05	-0.03	-0.03	-0.01	-0.1
QD	32	787	-0.02	-	-0.05	-0.05	0	-0.05
QE	30	1232	-0.03	-0.03	-0.02	-0.14	0	0
Curricular Branch:								
Scientific-Humanistic	32	9093	0	0	0.01	0	0	0.01
Technical-Professional	30	532	0.02	0.02	-	-0.03	0.03	-0.21
High School Type:								
Municipal	32	2098	-0.08	-0.1	-0.11	-0.13	-0.04	-0.15
Subsidized	32	3670	0.03	0.01	0.01	0.02	0	-0.04
Private	31	3857	0.02	0.06	0.06	0.1	0.11	0.21
Region:								
Center	31	5494	0.04	0.01	0	0.02	0.1	0.07
North	29	556	-0.02	-0.06	-0.07	0.02	-0.11	-0.07
South	32	3575	-0.04	-0.05	-0.05	-0.04	-0.03	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 478: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ciencias_Sociales_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	30	3985	-0.13	-0.12	-0.13	-0.19	-0.06	-0.08
Female	30	3992	0.15	0.13	0.15	0.2	0.06	0.07
SES:								
QA	30	757	0.05	0.03	0.02	0	-0.01	-0.01
QB	30	874	-0.04	-0.07	-0.08	0.03	-0.07	-0.1
QC	30	859	-0.07	-0.06	-0.08	0.12	-0.08	-0.09
QD	30	761	0	-	0.01	-0.1	0.02	-0.03
QE	30	1129	0.09	0.09	0.09	-0.19	0.13	0.12
Curricular Branch:								
Scientific-Humanistic	30	7315	0	0.01	0	-0.02	0.01	0.01
Technical-Professional	30	661	-0.13	-0.18	-0.18	0.27	-0.23	-0.4
High School Type:								
Municipal	30	2221	-0.07	-0.08	-0.08	-0.03	-0.04	-0.09
Subsidized	30	3299	0.04	0.02	0.03	-0.01	0	-0.04
Private	30	2456	0.02	0.07	0.07	0.21	0.12	0.16
Region:								
Center	30	5825	0.01	0.01	-0.01	-0.02	0.03	0.01
North	29	402	-0.03	-0.04	-0.03	0.45	-0.11	-0.12
South	30	1750	0.02	0.01	-	0.16	-0.02	-0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 479: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ciencias_Sociales_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	20	636	-0.39	-0.41	-0.42	-0.53	-0.29	-0.29
Female	20	3018	0.1	0.1	0.1	0.11	0.08	0.07
SES:								
QA	20	569	0.06	0.03	0.02	0.01	-0.05	-0.06
QB	20	691	0.19	0.17	0.18	-0.01	0.13	-0.04
QC	20	558	-0.01	-0.02	-0.01	0.07	0.01	-0.05
QD	20	346	-0.07	-0.04	-0.05	-0.04	0.05	0.07
QE	20	353	0.08	0.08	0.1	0.29	0.15	0.3
Curricular Branch:								
Scientific-Humanistic	20	3118	-0.01	-0.01	-0.01	-0.03	0.01	0.04
Technical-Professional	20	536	0.1	0.1	0.07	0.14	0.02	-0.17
High School Type:								
Municipal	20	1509	-	-0.01	-0.03	-0.12	-0.04	-0.11
Subsidized	20	1769	0.03	0.04	0.05	0.08	0.06	0.1
Private	19	376	-0.12	-0.1	-0.09	-0.12	0.14	0.25
Region:								
Center	19	1607	0.15	0.07	0.08	0.17	0.17	0.18
North	16	299	0.06	0.12	0.12	-0.49	-0.03	-0.01
South	20	1748	-0.02	0	-0.01	0.1	0.03	-0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 480: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Comunicaciones)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	2	224	-0.08	-0.08	-0.09	0.15	-0.06	-0.09
Female	2	184	0.13	0.14	0.16	-0.12	0.11	0.14
SES:								
QA	2	42	-0.25	-0.28	-0.22	-0.39	-0.23	-0.37
QB	2	32	0.13	0.08	0.14	-0.25	0.06	0.01
QC	2	44	-0.06	-0.09	-0.08	0.36	-0.1	-0.16
QD	2	42	0.07	0.07	0.01	-0.26	0.13	0.02
QE	2	55	0.17	0.18	0.14	0.36	0.09	0.21
Curricular Branch:								
Scientific-Humanistic	2	382	-	0	-	-0.05	-0.01	-
Technical-Professional	2	25	-0.01	-0.03	-0.03	0.32	0.02	-0.1
High School Type:								
Municipal	2	90	-0.02	-0.02	0.04	0.07	-0.01	-0.06
Subsidized	2	204	0.04	0.03	0.01	-0.19	0.01	-0.02
Private	2	113	0.03	0.05	0.05	0.14	0.05	0.14
Region:								
Center	2	279	0.06	0.05	0.04	0.02	0.08	0.06
North	2	47	-0.11	-0.11	-0.15	0.04	-0.2	-0.11
South	2	82	-0.15	-0.12	-0.11	-0.02	-0.19	-0.19

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 481: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Construcción)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	15	3646	-0.04	-0.02	-0.02	-0.02	-0.01	-0.03
Female	15	1107	0.16	0.11	0.08	0.12	0.05	0.1
SES:								
QA	15	602	-0.04	-0.07	-0.22	-0.03	-0.09	-0.1
QB	15	743	-0.07	-0.11	-0.03	-0.11	-0.14	-0.19
QC	15	724	0.05	0.03	0.01	0.06	0.05	0.01
QD	15	499	0.08	0.09	0.11	0.05	0.09	0.03
QE	15	531	0.07	0.1	0.17	0.07	0.15	0.25
Curricular Branch:								
Scientific-Humanistic	15	4292	-	0.02	0.03	0.01	0.02	0.02
Technical-Professional	15	461	0.03	-0.1	-0.08	-0.09	-0.08	-0.18
High School Type:								
Municipal	15	1641	-	-0.03	-0.11	-0.05	-0.05	-0.07
Subsidized	15	2407	0.02	0.02	0.04	0.03	0.04	0
Private	15	705	-0.05	0.02	0.07	0.01	0.1	0.18
Region:								
Center	15	2375	0.02	0.01	0.06	-0.02	0.14	0.01
North	13	270	-0.25	-0.21	-0.04	-0.27	-0.26	-0.3
South	15	2108	-	0.02	-0.02	0.02	-0.02	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 482: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Derecho)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	20	6766	-0.08	-0.07	-0.1	-0.15	0.01	-0.02
Female	20	5927	0.09	0.08	0.12	0.14	-0.02	0.01
SES:								
QA	20	1234	0.03	0.02	0.02	0.13	-0.02	-
QB	20	1043	0.03	0.01	-0.01	0.03	0	-0.04
QC	20	1123	-0.01	-0.03	-0.05	-0.03	-0.02	-0.08
QD	20	995	-0.08	-0.1	-0.09	-0.1	-0.06	-0.13
QE	20	1667	-	-0.01	-0.01	0.1	0.02	0.02
Curricular Branch:								
Scientific-Humanistic	20	12108	-0.01	-	-	0	-	0
Technical-Professional	20	584	-0.05	-0.1	-0.12	0.02	-0.14	-0.37
High School Type:								
Municipal	20	2682	-	-0.04	-0.04	-0.1	-0.02	-0.11
Subsidized	20	4807	-0.01	-0.02	-0.03	0.07	-0.05	-0.11
Private	20	5203	-0.04	-0.03	-0.02	-0.03	0.06	0.15
Region:								
Center	20	7181	0.07	0.04	0.02	0.02	0.12	0.01
North	20	1120	-0.18	-0.18	-0.15	-	-0.29	-0.37
South	20	4391	-0.04	-0.05	-0.06	-0.15	-0.07	-0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 483: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Diseño)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	19	2006	-0.1	-0.09	-0.09	-0.08	-0.01	-0.08
Female	19	3595	0.07	0.06	0.06	0.06	0.02	0.07
SES:								
QA	19	503	0.05	0.06	0.06	-0.04	0.06	0.02
QB	18	472	0	-0.02	-0.07	-0.23	-0.02	-0.06
QC	19	542	0.01	0.03	-0.09	-0.01	-0.01	-0.05
QD	19	465	0.03	-	0.01	-0.07	0	0.04
QE	19	789	-0.03	-0.01	0.01	0.15	0.01	0.01
Curricular Branch:								
Scientific-Humanistic	19	5232	-0.02	-0.02	-0.02	0.01	-0.01	0
Technical-Professional	17	368	0.08	0.05	0.09	-0.07	0.01	-0.19
High School Type:								
Municipal	18	1050	0.02	-	-0.03	0.04	0.01	-0.04
Subsidized	19	2109	0.01	-0.01	0.01	-0.08	-0.02	-0.04
Private	19	2441	-0.09	-0.07	-0.05	0.04	0	0.06
Region:								
Center	19	4355	-0.06	-0.06	-0.09	0.1	-0.04	0.03
North	17	272	-0.12	-0.12	-0.21	0.13	-0.27	-0.25
South	19	973	0.06	0.05	0.1	-0.09	0.04	0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 484: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Educación)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	68	4615	-0.16	-0.14	-0.14	-0.19	-0.06	-0.09
Female	68	6870	0.12	0.1	0.1	0.13	0.04	0.06
SES:								
QA	68	1745	0	0.01	-	-0.23	-0.03	-0.06
QB	68	2538	-0.05	-0.06	-0.07	0.15	-0.08	-0.11
QC	68	1722	0.01	0.01	0.01	-0.11	0.01	0.01
QD	68	1206	-0.07	-0.07	-0.08	-0.06	-0.05	-0.02
QE	63	911	0.09	0.1	0.11	-0.17	0.16	0.2
Curricular Branch:								
Scientific-Humanistic	68	9369	-0.01	-0.01	-	0.01	-0.01	0.02
Technical-Professional	68	2116	0.05	0.04	0.02	-0.12	0.02	-0.1
High School Type:								
Municipal	68	4927	-0.03	-0.03	-0.04	-0.03	-0.04	-0.09
Subsidized	68	5981	0.02	0.01	0.02	0.02	0.02	0.05
Private	65	577	0.04	0.05	0.1	-0.03	0.2	0.36
Region:								
Center	65	5630	0.01	0.01	-	0.18	0.07	0.01
North	56	523	-0.16	-0.16	-0.16	-0.75	-0.13	-0.06
South	66	5332	-0.05	-0.05	-0.05	0.05	-0.08	-0.09

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 485: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Educación_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	47	967	-0.5	-0.52	-0.47	-0.52	-0.31	-0.36
Female	59	11200	0.05	0.05	0.04	0.05	0.03	0.04
SES:								
QA	59	2102	0.02	0.01	-0.03	0.09	-0.05	-0.07
QB	59	2655	-0.02	-0.01	-0.01	-0.03	-0.03	-0.04
QC	59	1682	0	-	0.02	0.04	0.03	0.03
QD	59	1013	-0.1	-0.1	-0.11	-0.01	-0.03	-0.02
QE	58	947	-0.04	-0.06	-	-0.04	0.04	0.05
Curricular Branch:								
Scientific-Humanistic	59	9623	-0.03	-0.03	-0.03	-0.01	-0.01	0.02
Technical-Professional	59	2544	0.08	0.05	0.03	0.06	-	-0.09
High School Type:								
Municipal	59	5444	-0.03	-0.05	-0.05	-0.03	-0.08	-0.11
Subsidized	59	5502	0.01	0.01	0.01	0	0.05	0.06
Private	53	1221	-0.11	-0.07	-0.05	0.08	0.14	0.46
Region:								
Center	57	5259	-0.04	-0.05	-0.06	0.02	0.01	-0.02
North	42	566	-0.08	-0.07	0.06	-0.46	-0.13	-0.15
South	58	6342	-0.05	-0.05	-0.03	-0.13	-0.07	-0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 486: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Educación_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	35	2375	-0.14	-0.1	-0.2	-0.14	-0.06	-0.08
Female	35	3972	0.09	0.06	0.14	0.09	0.03	0.04
SES:								
QA	35	1192	-0.02	-0.03	0.09	0.02	-0.06	-0.02
QB	35	1523	0.02	0.02	-0.05	0.02	-	-0.02
QC	35	962	-0.02	-0.03	-0.04	-0.01	-	0.01
QD	35	561	-0.01	-0.02	-0.06	-0.07	0.01	-0.03
QE	35	385	0.02	0.01	0.07	0.01	0.08	0.09
Curricular Branch:								
Scientific-Humanistic	35	5554	-0.02	-0.01	0.01	-0.02	-0.01	0.01
Technical-Professional	35	793	0.04	-	-0.12	0.03	0.02	-0.1
High School Type:								
Municipal	35	3042	-0.01	-0.03	-	-0.01	-0.05	-0.06
Subsidized	35	3128	-	0.01	-0.04	-	0.03	0.04
Private	31	177	-0.2	-0.09	-0.12	-0.12	0.1	0.38
Region:								
Center	35	2624	-0.07	-0.02	-0.11	-0.07	0.11	0.09
North	24	252	-0.11	-0.1	-0.15	-0.12	-0.14	-0.1
South	35	3471	-0.07	-0.06	-0.04	0.03	-0.09	0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 487: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Educación_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	14	2531	-0.09	-0.09	-0.08	-0.07	-0.04	-0.05
Female	14	1762	0.16	0.15	0.16	0.14	0.09	0.12
SES:								
QA	18	553	-0.02	-0.05	-0.08	-0.02	-0.07	-0.09
QB	18	788	-0.02	-0.03	-0.03	0.07	-0.08	-0.11
QC	18	612	-0.01	-0.01	0.08	-	-0.02	-0.04
QD	18	471	-0.01	0.02	-0.05	-0.07	0.03	0.04
QE	18	455	-	0.03	-0.13	-0.06	0.1	0.13
Curricular Branch:								
Scientific-Humanistic	18	3773	-0.02	-0.02	-0.02	-0.02	-0.01	0.03
Technical-Professional	18	518	0.12	0.1	0.14	0.12	0.05	-0.1
High School Type:								
Municipal	18	1532	-0.01	-0.02	-0.02	-0.03	-0.07	-0.1
Subsidized	18	2212	0.02	0.02	0.04	0.03	0.05	0.07
Private	16	547	-0.05	0.01	-0.15	-0.02	0.08	0.25
Region:								
Center	17	1986	0.05	0.04	-0.1	0.03	0.08	0.07
North	12	286	-0.15	-0.17	-	-0.18	-0.2	-0.19
South	18	2020	0.04	0.03	0.04	0.06	0.02	0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 488: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (General)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	834	-0.06	-0.06	-0.09	-0.06	-0.01	-0.02
Female	1	879	0.06	0.05	0.08	0.07	0.01	0.02
SES:								
QA	1	107	-0.11	-0.12	0.03	-0.11	-0.12	-0.04
QB	1	135	-0.13	-0.12	0.01	-0.17	-0.13	-0.15
QC	1	146	0.09	0.08	0.21	0.04	0.08	0.02
QD	1	161	0.02	0.03	-0.12	0.07	0.04	0.02
QE	1	267	-0.11	-0.1	0	-0.08	-0.1	-0.1
Curricular Branch:								
Scientific-Humanistic	1	1682	0.01	0.01	0.02	0.01	0.01	0.01
Technical-Professional	1	31	-0.3	-0.27	-0.47	-0.32	-0.33	-0.33
High School Type:								
Municipal	1	454	0.02	0.01	0.03	0	0.08	0.06
Subsidized	1	611	-0.01	-0.01	-0.03	0.01	-0.06	-0.06
Private	1	648	0	0.01	0.01	-	0	0.01
Region:								
Center	1	1423	0.01	0.01	0	0.01	0.02	0.02
North	1	71	-0.42	-0.42	-0.25	-0.33	-0.47	-0.42
South	1	219	0.07	0.07	0.08	0.07	0.02	0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 489: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Humanidades)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	15	1269	-0.16	-0.13	-0.13	-0.25	-0.01	-0.04
Female	15	2049	0.1	0.08	0.09	0.13	-0.01	-
SES:								
QA	14	283	-0.02	0	-0.02	0.15	0.04	-
QB	15	336	-0.02	-0.02	0.01	0	-0.14	-0.08
QC	15	341	0.04	-0.02	-0.07	-0.51	-0.02	-0.06
QD	15	313	-0.05	-0.03	0.01	-	-0.1	-0.05
QE	15	438	0.02	-0.01	-0.05	0.21	0.11	-0.04
Curricular Branch:								
Scientific-Humanistic	15	3108	0.01	0.01	0.01	0	0.01	0.02
Technical-Professional	14	209	-0.01	-0.04	-0.07	-1.22	-0.09	-0.27
High School Type:								
Municipal	14	843	-0.04	-0.05	-0.08	-0.07	-0.12	-0.17
Subsidized	15	1479	-0.01	-0.01	0	-0.08	-	-0.03
Private	15	995	0.02	0.02	0.08	-0.2	0.09	0.2
Region:								
Center	15	2622	0.03	0.02	0.01	0.02	0.03	0.04
North	13	97	-0.19	-0.16	-0.16	-1.81	-0.09	-0.08
South	15	598	-0.03	-0.02	-0.03	0.24	-0.12	-0.17

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 490: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ingeniería_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	132	32640	-0.04	-0.02	-0.04	-0.04	0	0
Female	132	9619	0.14	0.07	0.07	0.13	-	0.01
SES:								
QA	131	4023	0.06	0.04	0.02	0.05	-	-0.01
QB	130	4540	-	-0.04	-0.01	-0.02	-0.08	-0.11
QC	131	4531	0	-0.03	0.04	-0.02	-0.04	-0.05
QD	131	3890	0.01	0.01	-0.05	0.01	0	-0.02
QE	130	5295	-0.01	0.02	0.08	0	0.06	0.08
Curricular Branch:								
Scientific-Humanistic	132	38585	-0.02	0	0	-0.01	0.01	0.03
Technical-Professional	130	3669	0.13	-0.01	0.01	0.08	-0.06	-0.15
High School Type:								
Municipal	132	11937	0.03	-	-0.05	-0.01	-0.04	-0.08
Subsidized	132	17937	0.02	0.01	0.04	0.03	0.01	0
Private	127	12380	-0.1	0.01	-0.03	-0.02	0.11	0.22
Region:								
Center	131	20512	-0.04	-0.05	-0.05	-0.07	0.03	-0.01
North	110	5020	0	-	-0.08	0	-0.02	-
South	125	16724	-0.02	-0.01	-0.04	-	-0.02	-0.02

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 491: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ingeniería_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	74	7886	-0.06	-0.04	-0.06	-0.05	-	-0.02
Female	73	4026	0.07	0.03	0.05	0.05	-	0.03
SES:								
QA	74	1398	0.01	-0.03	-0.02	-0.02	-0.07	-0.07
QB	73	2039	0	-0.02	-0.09	-0.03	-0.02	-0.03
QC	74	1837	-0.02	-0.04	-0.07	0.01	-0.02	-0.04
QD	73	1446	0.04	0.04	-0.07	0.04	0.05	0.04
QE	71	1245	-0.03	0.02	-0.06	0	0.06	0.01
Curricular Branch:								
Scientific-Humanistic	74	10015	-0.01	0	0.01	-0.01	0.01	0.02
Technical-Professional	74	1894	0.07	-0.01	-0.03	0.04	-0.01	-0.14
High School Type:								
Municipal	74	4350	0	-0.02	-0.04	-0.01	-0.02	-0.03
Subsidized	74	6501	0.01	0.01	0	0.02	0	-
Private	68	1058	-0.07	0.03	0.1	-0.01	0.15	0.21
Region:								
Center	68	7008	-0.02	-0.02	-0.11	-0.03	0.03	0.01
North	56	569	0.03	-0.01	0.2	0.05	-0.02	-
South	73	4333	-	-0.03	-0.04	0.02	-0.03	-0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 492: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Ingeniería_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	13	1204	-0.14	-0.09	-0.11	-0.16	-0.03	-0.06
Female	13	1192	0.12	0.09	0.06	0.14	0.01	0.04
SES:								
QA	13	192	0.08	0.11	0.24	0.08	0.04	0
QB	13	231	-0.09	-0.15	0.25	-0.1	-0.19	-0.18
QC	13	236	0.02	0.01	-0.57	0.01	0.03	-0.04
QD	13	211	-0.03	-0.02	-0.35	-0.07	0	0.02
QE	13	310	0.12	0.15	0.36	0.15	0.18	0.19
Curricular Branch:								
Scientific-Humanistic	13	2298	-0.01	-0.01	0.02	-0.01	-0.01	-
Technical-Professional	10	97	0.23	0.07	-0.19	0.15	-0.02	-0.09
High School Type:								
Municipal	13	717	-0.08	-0.08	0.36	-0.08	-0.07	-0.14
Subsidized	13	1153	0.06	0.05	-0.21	0.06	0.02	0.02
Private	13	525	-0.03	0.06	0.16	0.04	0.12	0.35
Region:								
Center	13	979	0.11	0.15	-0.01	0.06	0.14	0.09
North	13	230	0.01	-0.01	-0.2	-0.05	-0.09	-0.04
South	13	1187	0.02	-	-0.11	0	-0.02	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 493: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Mar)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	19	866	-0.11	-0.07	-0.02	-0.09	0.04	-0.05
Female	19	974	0.1	0.07	-0.12	0.08	-0.04	0.05
SES:								
QA	19	255	0.08	0.09	-0.04	0.01	-0.04	-0.1
QB	19	319	-0.03	-0.04	0.03	-0.05	-0.13	-0.15
QC	19	262	0.12	0.08	0.07	0.1	0.1	0.08
QD	18	170	0.11	0.04	0.03	0.06	0.07	0.15
QE	17	172	-0.01	0.06	0.08	-0.05	0.05	-
Curricular Branch:								
Scientific-Humanistic	19	1678	-	-	-0.03	-	0	0.02
Technical-Professional	19	162	0.07	0.05	0.21	0.12	0.05	-0.13
High School Type:								
Municipal	19	793	0.01	-0.01	-0.14	-0.02	-0.05	-0.05
Subsidized	19	839	0.01	0	0.08	-0.01	-0.01	-0.06
Private	17	208	-0.1	-0.04	0.04	-0.1	0.09	0.32
Region:								
Center	16	544	0.06	0.04	-0.36	0.01	0.11	-0.04
North	12	234	-0.09	-0.1	-0.42	-0.05	0.14	0.03
South	18	1062	0.1	0.09	0.09	0.02	0.05	0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 494: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Periodismo)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	16	1770	-0.17	-0.17	-0.19	-0.39	-0.08	-0.11
Female	16	2729	0.15	0.14	0.15	0.27	0.06	0.08
SES:								
QA	16	404	-0.08	-0.07	-0.07	-0.05	-0.11	-0.03
QB	15	347	0.01	-0.01	-0.05	-0.03	0.01	-0.12
QC	16	402	0.07	0.09	0.06	0.12	0.08	-0.03
QD	16	409	-0.05	-0.04	-0.06	-0.41	-	-0.05
QE	16	673	-0.01	0.02	0.02	0.3	-0.01	0.03
Curricular Branch:								
Scientific-Humanistic	16	4290	0	0	0	-0.01	0	0.01
Technical-Professional	14	209	-0.03	-0.06	-0.03	0.08	-0.13	-0.31
High School Type:								
Municipal	15	865	-0.03	-0.04	-0.04	-0.37	-0.03	-0.12
Subsidized	16	1738	0.02	-0.01	-0.01	0.16	-0.04	-0.07
Private	16	1896	-0.03	-0.01	-	0.21	0.03	0.05
Region:								
Center	16	2784	-0.08	-0.08	-0.08	-0.25	-0.06	-0.07
North	15	359	-0.04	-0.03	-0.02	0.23	-0.12	-0.06
South	16	1356	-0.02	-0.01	-0.02	-0.02	-0.08	-0.2

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 495: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Salud_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	31	5842	-0.1	-0.08	-0.08	-0.12	-0.01	-0.04
Female	31	6021	0.09	0.08	0.07	0.11	0.02	0.03
SES:								
QA	31	821	0.03	0.01	0.23	0.02	0.01	-0.01
QB	31	675	-0.11	-0.15	0.06	-0.12	-0.14	-0.19
QC	31	832	-0.12	-0.14	-0.15	-0.14	-0.12	-0.16
QD	31	918	0.04	0.04	0.01	0.04	0.01	-0.02
QE	31	1600	0.06	0.06	0.09	0.06	0.07	0.06
Curricular Branch:								
Scientific-Humanistic	31	11768	-	0	-	0	0	-
Technical-Professional	26	95	0.19	0.12	0.25	0.13	0.22	-0.08
High School Type:								
Municipal	31	2159	0.01	0.01	0.16	-0.01	0.04	-0.06
Subsidized	31	4483	-0.02	-0.03	0	-0.01	-0.05	-0.1
Private	31	5221	0.02	0.03	0.08	0.01	0.06	0.12
Region:								
Center	31	5582	-0.06	-0.09	0.09	-0.09	-0.03	-0.08
North	31	1471	-0.13	-0.14	-0.59	-0.13	-0.13	-0.07
South	31	4810	-0.06	-0.06	0.09	-0.04	-0.08	-0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 496: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Salud_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	60	4548	-0.21	-0.18	-0.08	-0.23	-0.1	-0.15
Female	60	13119	0.08	0.07	0.03	0.08	0.04	0.05
SES:								
QA	60	1741	0.01	0.01	0.07	0.02	-0.05	-0.07
QB	60	2515	-0.01	-0.03	-0.04	0	-0.04	-0.08
QC	60	2451	-0.02	-0.04	-0.09	-0.02	-0.04	-0.07
QD	60	2114	0	-0.01	-0.04	0.01	-0.01	-0.03
QE	60	2269	-0.07	-0.07	-0.09	-0.09	-0.02	0.08
Curricular Branch:								
Scientific-Humanistic	60	16904	-0.01	-	-0.01	-0.01	-	0.01
Technical-Professional	58	761	0.04	-	0.04	0.02	-	-0.12
High School Type:								
Municipal	60	5685	-0.02	-0.03	-0.01	-0.03	-0.04	-0.06
Subsidized	60	9049	0.01	0.01	-0.01	0.02	0.01	0
Private	57	2931	-0.06	-0.03	0.12	-0.06	0.04	0.17
Region:								
Center	59	8379	0.09	0.06	-0.03	0.04	0.11	0.05
North	52	1283	-0.06	-0.06	-0.22	-0.01	-0.06	0
South	58	8004	-0.05	-0.04	0.04	-0.02	-0.05	-0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 497: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Salud_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	26	2223	-0.18	-0.17	-0.25	-0.23	-0.09	-0.13
Female	26	5692	0.06	0.05	0.1	0.07	0.02	0.03
SES:								
QA	26	775	0.08	0.06	0.13	0.07	0.03	0.04
QB	26	1179	-0.03	-0.02	-0.07	-0.02	-0.05	-0.1
QC	25	1151	-0.01	-0.02	0.08	-0.04	-0.02	-0.01
QD	26	932	0	0.03	0.19	0.01	0.07	0.04
QE	26	925	-0.05	-0.04	-0.04	-0.04	0	-0.03
Curricular Branch:								
Scientific-Humanistic	26	7540	-0.01	-	-	-	0	0
Technical-Professional	25	372	0.09	0.03	0.08	0.05	0.01	-0.15
High School Type:								
Municipal	25	2621	-0.01	-0.02	-0.01	-0.01	-0.03	-0.07
Subsidized	26	4304	0.01	0	-0.08	0.02	0.02	0.02
Private	26	987	-0.08	-0.04	-0.17	-0.1	0.02	0.08
Region:								
Center	24	3963	0.02	0.01	-	0.01	0.08	0.06
North	24	512	-0.28	-0.25	-0.04	-0.31	-0.24	-0.07
South	25	3438	-0.01	-0.04	-0.11	-	-0.03	-0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 498: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Técnico_Administración)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	8	771	-0.16	-0.16	-0.15	0.06	-0.07	-0.05
Female	8	1127	0.08	0.08	0.08	0.09	0.02	0.02
SES:								
QA	8	241	-0.07	-0.08	-0.07	-0.25	-0.16	-0.25
QB	8	332	0.1	0.06	0.08	0.21	0.01	-0.07
QC	8	302	-0.04	-0.05	0	-0.12	-0.01	0.01
QD	8	246	-0.09	-0.02	-0.12	0.07	-0.03	-
QE	7	211	0	0.01	0.02	0.1	0.1	0.13
Curricular Branch:								
Scientific-Humanistic	8	1352	-0.07	-0.06	-0.06	-0.02	-0.03	0.05
Technical-Professional	8	546	0.16	0.13	0.12	0.07	0.04	-0.06
High School Type:								
Municipal	8	639	0.09	0.07	0.09	0.02	0.05	-0.03
Subsidized	8	1103	-0.02	-0.02	-0.04	0.02	-0.02	0.03
Private	6	156	-0.34	-0.29	-0.2	-0.08	-0.11	0.11
Region:								
Center	8	1527	-0.04	-0.02	-0.02	-0.01	-0.02	0.07
North	7	180	-0.17	-0.18	-0.22	-0.21	-0.15	0.03
South	8	191	-0.02	-0.04	-0.07	0.19	-0.1	-0.12

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 499: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Técnico_Agro)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	10	-0.13	-0.11	0.11	-0.11	-0.03	0.05
Female	1	5	0.35	0.29	-0.23	0.3	0.07	-0.13
SES:								
QA	1	3	0.08	0.15	-0.41	0.19	-0.02	-0.15
QB	1	4	-0.07	-0.22	-0.41	-0.23	-0.14	-0.03
QC	1	5	-0.34	-0.21	0.68	-0.22	-0.19	-0.32
QD	1	1	0.44	0.5	0.59	0.51	0.95	0.92
QE	-	-	-	-	-	-	-	-
Curricular Branch:								
Scientific-Humanistic	1	10	-0.1	-0.06	-0.03	-0.06	-0.03	0.04
Technical-Professional	1	5	0.18	0.11	0.07	0.12	0.05	-0.1
High School Type:								
Municipal	1	9	-0.07	-0.08	-0.11	-0.06	-0.1	-0.32
Subsidized	1	6	0.16	0.16	0.56	0.13	0.19	0.46
Private	-	-	-	-	-	-	-	-
Region:								
Center	1	1	0.81	0.71	-	0.67	0.67	0.92
North	-	-	-	-	-	-	-	-
South	1	14	-0.05	-0.04	0.01	-0.04	-0.05	-0.08

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 500: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Técnico_Ciencias)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	5	238	-0.16	0.08	-0.14	-0.1	-0.06	-0.03
Female	5	318	0.07	0.01	0.11	0.09	0.03	0.01
SES:								
QA	5	63	0.25	0.18	0.06	0.1	0.22	0.14
QB	5	124	-0.15	-0.12	0.02	-0.07	-0.22	-0.1
QC	5	70	-0.07	-0.05	-0.89	-0.05	-0.09	-0.16
QD	5	56	0.1	-0.12	0.23	0.26	0.19	0.28
QE	4	37	-0.22	-0.09	0.19	-0.17	-0.02	0.01
Curricular Branch:								
Scientific-Humanistic	5	497	-0.04	-0.05	0.02	-0.02	-0.03	-0.01
Technical-Professional	5	59	0.09	0.08	-0.17	0.02	0.06	-0.06
High School Type:								
Municipal	5	313	0.03	0.03	0.02	-	-0.01	-0.02
Subsidized	5	230	-0.04	-0.13	-0.06	0.01	0.03	0.03
Private	4	13	0.3	0.42	0.21	0.15	0.56	0.61
Region:								
Center	4	181	0.7	0.71	-	0.66	0.8	0.61
North	4	9	-0.82	-0.67	-	-0.98	-0.76	-1.05
South	4	366	0	0	-0.01	0	-0.09	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 501: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Técnico_Diseño)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	3	211	-0.08	-0.09	-0.05	-0.02	-0.07	-0.05
Female	3	155	0.19	0.23	0.16	0.03	0.2	0.22
SES:								
QA	3	39	-0.29	-0.34	-0.44	-0.58	-0.38	-0.3
QB	3	74	0.19	0.12	0.25	-0.26	0.07	0.07
QC	3	59	0.17	0.2	-0.1	-0.13	0.24	0.23
QD	3	49	-0.06	-0.06	0.08	0.16	-0.03	-0.03
QE	3	35	0.13	0.17	0.46	0.05	0.25	0.28
Curricular Branch:								
Scientific-Humanistic	3	288	-0.06	-0.03	-0.07	-0.05	0	0.03
Technical-Professional	3	78	0.15	0.05	0.15	0.31	-0.04	-0.1
High School Type:								
Municipal	3	124	0.09	0.06	0.09	-0.17	0.01	-0.03
Subsidized	3	226	-0.03	-0.02	-0.04	0.03	-0.01	0.02
Private	3	16	-0.16	-0.14	-0.2	0.89	-0.01	-0.24
Region:								
Center	3	346	0.01	0.01	-0.05	-0.01	0.01	0.02
North	2	3	-0.02	0.01	1.53	-0.46	-0.05	-0.99
South	3	17	-0.04	-0.1	0	2.41	-0.05	0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 502: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Técnico_Educación)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	84	-0.03	-0.01	0.04	0.03	0.01	-0.02
Female	1	52	0.06	0.02	-0.12	-0.03	-0.02	0.03
SES:								
QA	1	16	0.14	0.13	0.16	0.2	0.12	0.03
QB	1	29	-0.03	-0.03	-0.05	-0.04	-0.07	-0.17
QC	1	26	-0.06	-0.06	-0.1	-0.22	-0.03	0.1
QD	1	18	-0.56	-0.55	-0.72	-0.35	-0.5	-0.51
QE	1	17	0.17	0.21	0.14	0.25	0.15	0.23
Curricular Branch:								
Scientific-Humanistic	1	118	0.03	0.03	0.02	0.07	0.03	0.05
Technical-Professional	1	18	-0.09	-0.1	0.01	-0.36	-0.08	-0.3
High School Type:								
Municipal	1	38	0.08	0.13	0.22	0.15	0.1	-0.02
Subsidized	1	91	-0.05	-0.07	-0.14	-0.09	-0.07	-0.03
Private	1	7	0.35	0.34	0.62	0.36	0.39	0.52
Region:								
Center	1	116	-0.04	-0.04	-0.05	-0.06	-0.04	-0.05
North	1	8	0.41	0.44	0.44	0.33	0.36	0.2
South	1	12	0.08	0.14	0.02	0.27	0.14	0.35

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 503: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Técnico_Idioma)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	6	276	-0.13	-0.12	-0.14	0.06	-0.05	-0.08
Female	6	590	0.07	0.07	0.08	-0.01	0.02	0.04
SES:								
QA	6	97	0.02	0	0	0.42	0.01	-0.04
QB	6	153	-0.05	-0.04	-0.06	-0.26	-0.07	-0.29
QC	6	127	-0.25	-0.25	-0.27	-0.45	-0.24	0.12
QD	6	98	0.14	0.05	0.16	0.09	0.05	-0.17
QE	5	74	0.11	0.07	0.15	0.04	0.14	-0.08
Curricular Branch:								
Scientific-Humanistic	6	758	-0.04	-0.03	-0.04	0.03	-0.04	0.03
Technical-Professional	6	107	0.19	0.16	0.15	-0.29	0.16	-0.06
High School Type:								
Municipal	6	321	0.03	0.02	0.01	0.01	0.01	-0.12
Subsidized	6	466	-0.04	-0.04	-0.03	-0.05	-0.05	0.02
Private	6	78	-0.01	0.08	-0.06	0.57	0.21	0.09
Region:								
Center	6	400	0.14	0.14	0.13	0.24	0.19	0.04
North	6	38	0.06	0.09	0.04	0.52	-0.04	0.35
South	6	428	-0.11	-0.06	-0.07	-0.19	-0.09	-0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 504: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Técnico_Ingeniería)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	64	7221	-0.04	-0.02	-0.02	-0.03	-0.02	-0.03
Female	62	2043	0.06	-0.02	0.03	0.01	-0.01	-
SES:								
QA	63	1170	-0.03	-0.05	-0.08	-0.03	-0.08	-0.07
QB	62	1832	0.03	0	-0.03	0	0.01	-0.01
QC	64	1521	0.02	0.01	-0.03	0.03	-	-0.03
QD	62	980	-0.01	0.01	0.04	-0.06	-	-0.01
QE	61	851	0.03	0.01	-0.02	0.02	0.09	0.07
Curricular Branch:								
Scientific-Humanistic	64	6821	-0.08	-0.05	-0.09	-0.05	-0.03	-
Technical-Professional	63	2441	0.13	0.05	0.07	0.08	0.04	-0.02
High School Type:								
Municipal	64	3782	-0.02	-0.03	-0.04	-0.02	-0.04	-0.08
Subsidized	64	4947	0.04	0.05	0.04	0.07	0.06	0.05
Private	55	533	-0.14	-	-0.19	-0.08	0.13	0.34
Region:								
Center	61	6223	-0.08	-0.07	-0.15	-0.13	0	0.01
North	50	641	-0.13	-0.16	-0.17	-0.08	-0.17	-0.24
South	60	2400	0.01	0.01	0.04	0.05	-	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Table 505: Average Over Prediction (-) and Under Prediction (+) of FYGPA by Predictor Measure and Subgroups (Veterinaria)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	5	1149	-0.05	-0.03	-0.11	-0.06	0.05	0
Female	5	1662	0.04	0.03	0.08	0.05	-0.02	0.01
SES:								
QA	5	246	-0.01	-0.01	-0.02	-	-0.02	-0.09
QB	5	322	-0.07	-0.07	0	-0.07	-0.09	-0.13
QC	5	328	0.01	0	-0.12	0.01	-	-0.02
QD	5	296	-	-0.03	0.1	-0.04	-0.02	-0.04
QE	5	328	0.04	0.05	-0.03	0.03	0.12	0.13
Curricular Branch:								
Scientific-Humanistic	5	2644	-	-	-0.02	-0.01	0	0.02
Technical-Professional	5	166	0.05	0.01	0.18	0.07	-0.02	-0.21
High School Type:								
Municipal	5	795	-0.05	-0.06	-0.04	-0.04	-0.07	-0.11
Subsidized	5	1377	0.03	0.02	-0.02	0.03	0	-
Private	5	638	-	0.03	0.14	-0.02	0.12	0.21
Region:								
Center	5	1284	0.01	0	-	-0.04	0.05	0.01
North	5	178	-0.02	-0.01	0.01	0.02	-0.05	-0.02
South	5	1348	-0.06	-0.06	-	-0.06	-0.04	-0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed FYGPA. Prediction equations are estimated within careers.)

Appendix W. Prediction Bias by the Type of Career – Second Year Grade Point Average (SYGPA)

Table 506: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Administración)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	36	6751	-0.14	-0.13	-0.15	-0.14	-0.05	-0.08
Female	36	5693	0.17	0.16	0.18	0.18	0.06	0.1
SES:								
QA	36	1421	0.11	0.09	-	0.13	0.03	-0.01
QB	36	1150	0.03	0.01	-	-0.02	-0.07	-0.11
QC	36	895	-0.07	-0.09	-	-0.07	-0.09	-0.12
QD	35	711	-0.14	-0.11	-	-0.15	-0.07	-0.02
QE	36	1426	-0.02	0.01	-	0.08	0.08	0.08
Curricular Branch:								
Scientific-Humanistic	36	10908	-0.07	-0.05	-0.05	-0.03	-0.03	0.01
Technical-Professional	35	1536	0.23	0.18	0.14	0.18	0.14	0
High School Type:								
Municipal	36	2734	0.08	0.06	0.07	0.06	0.07	-0.06
Subsidized	36	3800	-0.01	-0.01	-0.01	-0.03	-0.03	-0.02
Private	35	5910	-0.25	-0.18	-0.14	-0.23	-0.05	0.16
Region:								
Center	33	7690	0.01	0.03	0.04	-0.01	0.08	-0.04
North	25	755	-0.36	-0.36	-0.31	-0.34	-0.41	-0.34
South	36	3999	0.04	0.03	0.05	0.04	0.03	0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 507: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Administración_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	26	2137	-0.14	-0.11	-0.15	-0.1	0	-0.07
Female	26	2646	0.12	0.09	0.12	0.08	0.01	0.01
SES:								
QA	25	871	0.12	0.13	-	0.05	0.05	-0.05
QB	25	1088	0.04	0.03	-	0.11	0.01	0.02
QC	24	609	-0.05	-0.05	-	-0.01	0	-0.1
QD	21	343	-0.04	-0.01	-	-0.09	0.08	0.14
QE	21	312	-0.31	-0.26	-	-0.21	-0.19	-0.01
Curricular Branch:								
Scientific-Humanistic	26	2611	-0.22	-0.18	-0.17	-0.08	-0.11	-0.16
Technical-Professional	26	2169	0.21	0.17	0.16	0.19	0.13	0.04
High School Type:								
Municipal	26	2390	0.03	0.03	0.03	0.05	0.05	-0.03
Subsidized	26	2069	-0.04	-0.04	-0.05	-0.04	-0.04	-0.05
Private	17	321	-0.35	-0.26	-0.37	-0.09	-0.03	0.22
Region:								
Center	21	2427	-0.04	-0.05	-0.06	-0.27	0.03	0.01
North	13	139	-0.38	-0.17	-0.3	0.1	-0.17	-0.22
South	25	2214	-0.05	0.02	0.06	0.01	0.03	0

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 508: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Administración_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	10	511	-0.23	-0.21	-0.2	-0.09	-0.08	-0.01
Female	10	960	0.15	0.14	0.13	0.1	0.07	-0.01
SES:								
QA	10	159	0.08	0.1	-	-0.15	0.03	0.01
QB	9	195	-0.02	-0.06	-	0.04	-0.09	-0.14
QC	9	176	0.05	-0.02	-	-0.02	-0.03	-0.31
QD	9	120	-0.08	-0.07	-	-0.01	-0.11	0.05
QE	10	186	-0.05	0	-	0.1	0.01	0.08
Curricular Branch:								
Scientific-Humanistic	10	1231	-0.03	-0.01	0	-0.01	0.05	0.06
Technical-Professional	9	238	0.17	0.09	0.05	0.08	-0.06	-0.17
High School Type:								
Municipal	10	374	0.06	0.14	0.01	0.09	0.03	-0.01
Subsidized	10	761	0.03	0.01	0.03	0	-0.01	-0.08
Private	10	334	-0.13	-0.07	-0.09	-0.01	0.03	0.21
Region:								
Center	9	1063	-0.02	-0.02	-0.04	-0.06	0.03	-0.05
North	9	111	0.05	0.15	0.14	-0.07	-0.01	0.12
South	9	295	0.08	0.06	0.09	0.29	0.06	0.29

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 509: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Agro)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	16	2585	-0.05	-0.03	-0.01	-0.04	0.02	0.01
Female	16	1977	0.06	0.03	0.01	0.05	-0.04	-0.02
SES:								
QA	16	545	0.06	0.06	-	0.07	-0.01	-0.07
QB	16	601	0.05	0.03	-	0.04	0.02	-0.03
QC	16	507	-0.11	-0.12	-	-0.11	-0.13	-0.12
QD	16	392	0.06	0.05	-	0.03	0.05	0
QE	15	508	-0.08	-0.06	-	-0.08	-0.01	0.06
Curricular Branch:								
Scientific-Humanistic	16	4176	-0.02	-0.01	-0.01	-0.01	-0.01	0.01
Technical-Professional	16	386	0.13	0.08	0.11	0.09	0.07	-0.09
High School Type:								
Municipal	16	1533	0.09	0.04	0.03	0.07	0.05	-0.03
Subsidized	16	1805	-0.05	-0.04	-0.05	-0.05	-0.06	-0.05
Private	14	1224	0	0.11	0.04	0.03	0.19	0.29
Region:								
Center	16	2459	-0.03	-0.05	-0.04	-0.02	0.01	0.11
North	14	126	-0.21	-0.21	-0.17	-0.23	-0.26	-0.23
South	16	1977	-0.03	-0.02	-0.07	-0.02	-0.04	-0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 510: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Agro_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	4	165	-0.06	-0.03	0.12	-0.07	0	-0.04
Female	4	230	0.1	0.06	-0.05	0.11	0.03	0.06
SES:								
QA	4	62	0.16	0.12	-	-0.04	0.1	-0.06
QB	4	96	-0.08	-0.06	-	-0.07	-0.12	0
QC	4	54	-0.42	-0.37	-	-0.29	-0.33	-0.35
QD	4	35	0.18	0.21	-	0.2	0.19	0.13
QE	4	34	-0.15	-0.16	-	-0.16	0.07	0.11
Curricular Branch:								
Scientific-Humanistic	4	330	-0.07	-0.02	0.11	-0.04	-0.03	0.01
Technical-Professional	4	65	0.22	0.09	-0.29	0.08	0.06	-0.04
High School Type:								
Municipal	4	139	0	-0.02	0.19	-0.01	-0.01	-0.08
Subsidized	4	217	-0.02	-0.01	-0.21	-0.03	-0.04	0.02
Private	3	39	0.26	0.1	0.19	0.25	0.29	0.25
Region:								
Center	4	278	-0.68	-0.73	0.07	-0.72	-0.66	-0.21
North	3	6	0.32	0.36	0.11	0.51	0.25	0.47
South	4	111	0.18	0.12	0.36	0.39	0.16	0.08

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 511: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Arquitectura)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	17	2186	-0.07	-0.06	-0.11	-0.1	-0.02	-0.04
Female	17	2192	0.09	0.07	0.1	0.11	0.02	0.03
SES:								
QA	15	392	0.04	0.14	-	0.06	0.13	0.08
QB	17	289	0.03	-0.04	-	0.03	-0.05	-0.18
QC	17	314	-0.08	-0.1	-	-0.16	-0.08	-0.16
QD	17	301	-0.14	-0.19	-	-0.13	-0.15	-0.11
QE	15	496	-0.15	-0.14	-	-0.11	-0.16	-0.13
Curricular Branch:								
Scientific-Humanistic	17	4141	0.01	0.01	0	0	0.01	0.01
Technical-Professional	16	237	-0.07	-0.09	-0.04	0.05	-0.11	-0.32
High School Type:								
Municipal	17	851	-0.1	-0.13	-0.09	-0.12	-0.13	-0.2
Subsidized	17	1456	-0.01	-0.02	-0.03	-0.01	-0.05	-0.08
Private	17	2071	-0.08	0.01	0	0.03	0.12	0.3
Region:								
Center	17	2411	0.11	0.04	0.02	0.19	0.14	0.14
North	15	372	-0.42	-0.42	-0.44	-0.04	-0.53	-0.44
South	17	1595	-0.03	0.01	-0.04	0	-0.02	0

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 512: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Arte_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	14	351	-0.16	-0.12	-0.19	-0.31	-0.05	-0.07
Female	14	1426	0.05	0.04	0.06	0.11	0.04	0.06
SES:								
QA	13	163	0.08	0.06	-	-0.04	0.1	0.13
QB	12	118	-0.08	-0.09	-	-0.1	-0.07	-0.12
QC	11	95	-0.24	-0.23	-	-0.31	-0.26	-0.24
QD	13	98	0.14	0.05	-	0.78	0.01	-0.03
QE	13	221	0.03	0.05	-	-0.03	0.02	-0.02
Curricular Branch:								
Scientific-Humanistic	14	1728	0.01	0	0	0.01	0.01	0.02
Technical-Professional	11	49	-0.02	0.06	-0.02	-0.01	0.03	-0.22
High School Type:								
Municipal	14	289	0	-0.02	-0.09	0.13	0.01	-0.1
Subsidized	14	549	-0.05	-0.09	-0.08	-0.19	-0.09	-0.11
Private	13	939	0.08	0.07	0.08	0.32	0.12	0.15
Region:								
Center	14	1284	0.03	-0.02	0	-0.5	0.02	0.02
North	12	58	0.18	0.23	0.22	0.02	0.18	0.24
South	14	435	-0.34	-0.4	-0.49	0.01	-0.2	-0.52

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 513: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Arte_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	11	472	-0.07	-0.03	-0.06	-0.11	-0.01	-0.03
Female	11	567	0.07	0.03	0.08	0.03	0	0.01
SES:								
QA	11	73	0.05	0.01	-	0.23	0.03	-0.05
QB	11	76	-0.13	-0.1	-	-0.33	-0.12	-0.21
QC	11	77	-0.09	-0.11	-	-0.22	0.01	0.07
QD	11	67	-0.2	-0.22	-	-0.11	-0.18	-0.24
QE	10	140	0	0	-	0.07	-0.01	0
Curricular Branch:								
Scientific-Humanistic	11	996	0.02	0.03	0.03	-0.02	0.02	0.01
Technical-Professional	10	43	-0.46	-0.44	-0.63	0.06	-0.34	-0.15
High School Type:								
Municipal	11	169	0.02	0.03	0.02	0.08	0.1	0.08
Subsidized	11	355	0	-0.02	-0.05	0.04	-0.09	-0.13
Private	11	515	0.11	0.12	0.07	-0.08	0.13	0.21
Region:								
Center	11	816	-0.06	-0.07	-0.09	-0.08	0.01	-0.02
North	8	45	-0.1	-0.12	-0.16	-0.25	-0.23	-0.22
South	11	178	0.07	0.06	0.06	0.48	0	0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 514: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ciencias)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	16	1284	-0.13	-0.09	-0.37	-0.13	-0.05	-0.05
Female	16	1617	0.12	0.09	0.21	0.12	0.05	0.1
SES:								
QA	16	209	-0.22	-0.2	-	-0.19	-0.22	-0.22
QB	16	346	-0.1	-0.12	-	-0.13	-0.18	-0.21
QC	15	287	0.08	0.08	-	0.05	0.2	0.08
QD	16	239	-0.01	0.07	-	0.06	-0.01	-0.04
QE	15	269	0.08	0.11	-	0.11	0.1	0.09
Curricular Branch:								
Scientific-Humanistic	16	2721	0	0.01	0.03	0.01	0.01	0.04
Technical-Professional	16	180	0.01	-0.12	-0.48	-0.1	0.02	-0.24
High School Type:								
Municipal	16	832	-0.03	-0.06	-0.2	-0.07	-0.01	-0.06
Subsidized	16	1276	0.05	0.05	0.27	0.05	0.02	0.04
Private	16	793	0.15	0.22	-0.04	0.17	0.21	0.22
Region:								
Center	15	2147	0.03	0.04	-0.54	0.05	0.07	0.08
North	15	151	0.24	0.22	0.39	0.21	0.12	0.16
South	16	603	-0.12	-0.12	0.12	-0.14	-0.23	-0.15

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 515: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ciencias_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	26	2315	-0.05	-0.03	-0.13	-0.07	0.02	-0.01
Female	26	2916	0.02	0	0.06	0.03	-0.04	0.02
SES:								
QA	25	481	0.01	0.03	-	0.01	0	0.01
QB	26	798	0.02	-0.01	-	0	-0.02	-0.08
QC	26	677	-0.03	-0.04	-	-0.02	-0.05	-0.07
QD	26	517	0.03	0.04	-	0.02	0.06	0.02
QE	26	522	0.04	0.07	-	0.06	0.08	0.1
Curricular Branch:								
Scientific-Humanistic	26	5023	0	0	0.01	0	0	0.01
Technical-Professional	26	207	0.02	-0.06	-0.38	-0.03	-0.04	-0.17
High School Type:								
Municipal	26	1840	-0.01	-0.01	-0.2	-0.02	-0.01	-0.05
Subsidized	26	2589	0.02	0	0.09	0.02	0	-0.02
Private	25	801	-0.1	-0.06	0.08	-0.1	0.01	0.23
Region:								
Center	26	2824	-0.06	-0.11	0.15	-0.11	-0.08	-0.01
North	25	535	0.1	0.13	-0.05	0.15	0.06	0.07
South	25	1871	-0.08	-0.08	-0.02	-0.07	-0.08	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 516: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ciencias_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	10	655	-0.03	-0.02	0.04	-0.03	0.03	0.02
Female	10	441	0.03	0.01	-0.09	0.03	-0.07	-0.07
SES:								
QA	10	133	-0.02	0.02	-	0	-0.02	-0.17
QB	10	172	-0.06	-0.09	-	-0.02	-0.09	-0.13
QC	10	122	-0.24	-0.2	-	-0.2	-0.19	-0.22
QD	10	88	-0.17	-0.2	-	-0.23	-0.14	-0.16
QE	10	97	-0.09	-0.12	-	-0.19	-0.08	-0.16
Curricular Branch:								
Scientific-Humanistic	10	1015	-0.01	0.01	0.01	0.01	0	0.01
Technical-Professional	9	80	0.15	-0.01	-0.11	-0.03	0.14	-0.22
High School Type:								
Municipal	10	367	0.06	0.03	0.05	0.04	0.05	0
Subsidized	10	501	0.03	0.02	0.08	0.01	0.01	-0.06
Private	10	227	-0.12	-0.04	-1.06	-0.18	0.02	0.2
Region:								
Center	10	595	0.06	0.05	-0.09	0.04	0.1	-0.08
North	10	96	-0.2	-0.2	0.14	-0.16	-0.22	-0.07
South	10	405	-0.11	-0.09	-0.17	-0.08	-0.12	-0.17

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 517: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ciencias_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	4	278	-0.04	0.03	0.01	-0.04	0.04	0.04
Female	4	168	0.06	-0.06	-0.29	0.06	-0.06	-0.08
SES:								
QA	4	21	0.08	0.07	-	0.05	-0.01	-0.28
QB	4	33	-0.28	-0.26	-	-0.16	-0.41	-0.36
QC	4	34	0.23	0.02	-	0.39	0.19	0.32
QD	4	27	0.19	0.22	-	0.16	0.18	0.4
QE	4	37	0.17	0.21	-	0.31	0.16	0.26
Curricular Branch:								
Scientific-Humanistic	4	425	-0.02	-0.01	0.03	-0.01	-0.01	-0.01
Technical-Professional	4	21	0.46	0.28	-0.63	0.31	0.27	0.13
High School Type:								
Municipal	4	136	0.06	0.02	-0.25	0.03	-0.07	-0.11
Subsidized	4	222	-0.05	-0.04	0.25	-0.03	0	0
Private	4	88	0.16	0.23	-0.45	0.15	0.26	0.27
Region:								
Center	4	41	0.37	0.41	0.94	0.24	0.42	0.27
North	4	154	-0.16	0.02	-0.03	-0.09	-0.18	-0.2
South	4	251	-0.23	-0.2	-0.35	-0.22	-0.16	-0.14

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 518: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ciencias_Sociales_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	31	2697	-0.26	-0.25	-0.24	-0.26	-0.15	-0.19
Female	31	4766	0.15	0.15	0.14	0.13	0.07	0.1
SES:								
QA	31	716	-0.02	0	-	-0.01	-0.03	-0.05
QB	31	690	0.01	-0.04	-	-0.06	-0.07	-0.12
QC	31	596	-0.08	-0.04	-	-0.06	-0.04	-0.13
QD	31	549	-0.05	-0.06	-	-0.15	-0.04	-0.02
QE	31	823	-0.03	-0.02	-	-0.18	-0.01	0.01
Curricular Branch:								
Scientific-Humanistic	31	7068	0	0	0	-0.01	0	0.02
Technical-Professional	29	394	-0.07	-0.07	-0.15	0.12	-0.1	-0.44
High School Type:								
Municipal	31	1606	-0.05	-0.06	-0.09	-0.06	-0.01	-0.15
Subsidized	31	2741	0	-0.02	-0.03	-0.08	-0.04	-0.07
Private	30	3115	0.01	0.05	0.06	0.14	0.1	0.19
Region:								
Center	30	4096	0.02	-0.03	-0.02	0.08	0.08	-0.01
North	27	533	-0.09	-0.06	-0.08	-0.35	-0.19	-0.06
South	31	2833	-0.05	-0.06	-0.07	-0.03	-0.09	-0.1

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 519: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ciencias_Sociales_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	30	2802	-0.14	-0.13	-0.14	-0.11	-0.05	-0.1
Female	30	2837	0.15	0.15	0.15	0.1	0.07	0.1
SES:								
QA	29	536	0.08	0.09	-	-0.18	0.09	0.02
QB	29	686	-0.13	-0.14	-	0.04	-0.13	-0.2
QC	29	620	-0.04	-0.05	-	0.05	-0.08	-0.1
QD	29	503	0.04	0.04	-	-0.08	0.07	0.03
QE	30	677	0.08	0.09	-	-0.18	0.12	0.11
Curricular Branch:								
Scientific-Humanistic	30	5165	-0.01	0	-0.01	-0.01	0	0
Technical-Professional	29	473	0.06	0.01	0.04	0.14	0	-0.05
High School Type:								
Municipal	30	1653	-0.07	-0.08	-0.08	0.1	-0.01	-0.1
Subsidized	30	2457	0.03	0.01	0.02	-0.12	-0.01	-0.03
Private	29	1528	0.01	0.04	0.06	0.37	0.1	0.18
Region:								
Center	30	4071	0.03	0.04	0.03	0	0.08	0.05
North	29	295	-0.02	-0.03	-0.02	0.02	-0.11	-0.15
South	29	1273	-0.03	-0.03	-0.03	0.06	-0.07	-0.1

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 520: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ciencias_Sociales_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	18	455	-0.45	-0.44	-0.45	-0.62	-0.32	-0.31
Female	18	2413	0.1	0.1	0.09	0.1	0.07	0.06
SES:								
QA	18	473	0.18	0.14	-	0.14	0.05	0.02
QB	18	588	0.02	0	-	-0.11	-0.03	-0.13
QC	18	405	-0.1	-0.09	-	-0.15	-0.06	-0.05
QD	18	231	-0.09	-0.05	-	-0.08	0.01	0.12
QE	18	209	-0.01	0.02	-	0.11	0.25	0.33
Curricular Branch:								
Scientific-Humanistic	18	2429	-0.04	-0.03	-0.03	-0.08	-0.03	0
Technical-Professional	18	439	0.22	0.16	0.16	0.31	0.15	-0.02
High School Type:								
Municipal	18	1223	0	-0.02	-0.02	-0.07	-0.03	-0.07
Subsidized	18	1359	0.04	0.05	0.05	0.13	0.05	0.07
Private	17	286	-0.11	-0.05	-0.07	-0.2	0.12	0.1
Region:								
Center	18	1168	0.1	0.09	0.09	0.01	0.2	0.01
North	13	240	-0.08	-0.06	-0.04	0.62	-0.07	0.11
South	18	1460	0.03	0.06	0.07	-0.26	0.05	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 521: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Comunicaciones)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	2	160	-0.06	-0.08	-0.07	0.03	-0.05	-0.08
Female	2	147	0.1	0.13	0.13	0	0.08	0.11
SES:								
QA	2	33	-0.03	-0.03	-	-0.23	0	-0.13
QB	2	22	0.15	0.17	-	-0.21	0.21	0.19
QC	2	33	-0.03	-0.05	-	0.46	-0.11	-0.23
QD	2	32	0.06	0.1	-	-0.03	0.17	0.1
QE	2	40	0.02	0.02	-	0.11	-0.01	-0.04
Curricular Branch:								
Scientific-Humanistic	2	290	-0.03	-0.03	-0.02	-0.09	-0.03	0
Technical-Professional	2	16	0.29	0.39	0.27	0.6	0.32	0.03
High School Type:								
Municipal	2	68	0	0	0.07	0.51	0.02	-0.11
Subsidized	2	149	0.02	0.03	0	-0.25	0.01	0
Private	2	89	0.03	0.03	-0.01	0.02	0.03	0.17
Region:								
Center	2	221	-0.01	-0.01	0.02	0.05	0	0
North	2	32	-0.13	-0.18	-0.29	-0.02	-0.2	-0.25
South	2	54	0.14	0.18	0.13	-0.22	0.15	0.17

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 522: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Construcción)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	15	2617	-0.03	-0.03	-0.04	-0.03	0	-0.02
Female	15	796	0.1	0.11	0.11	0.13	0.01	0.03
SES:								
QA	15	439	-0.01	-0.05	-	-0.05	-0.01	-0.07
QB	15	563	-0.1	-0.1	-	-0.07	-0.11	-0.09
QC	15	480	0.03	0.07	-	0.04	0.04	-0.14
QD	15	314	-0.06	-0.07	-	-0.04	-0.1	-0.1
QE	15	332	0.13	0.12	-	0.1	0.21	0.22
Curricular Branch:								
Scientific-Humanistic	15	3122	0	0.01	0	0.01	0.01	0.01
Technical-Professional	15	291	-0.07	-0.12	-0.06	-0.09	-0.13	-0.18
High School Type:								
Municipal	15	1199	-0.11	-0.12	-0.21	-0.12	-0.05	-0.24
Subsidized	15	1689	0.03	0.03	0.09	0.05	0.05	0.04
Private	15	525	0.05	0.09	0.14	0.04	0.15	0.29
Region:								
Center	14	1758	-0.07	-0.05	0.04	-0.07	0.02	-0.02
North	12	145	0.08	0.02	-0.11	0.01	-0.05	-0.2
South	15	1510	-0.06	-0.07	0.06	-0.1	-0.09	-0.12

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 523: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Derecho)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	20	5154	-0.09	-0.08	-0.1	-0.09	-0.01	-0.05
Female	20	4566	0.09	0.09	0.11	0.08	0.01	0.05
SES:								
QA	20	901	0.07	0.08	-	-0.06	0.04	0.04
QB	20	806	0.07	0.07	-	0.17	0.08	0
QC	20	806	-0.01	-0.04	-	-0.17	-0.04	-0.07
QD	20	688	-0.14	-0.16	-	-0.11	-0.12	-0.18
QE	20	1141	0.06	0.05	-	0.08	0.07	0.08
Curricular Branch:								
Scientific-Humanistic	20	9320	-0.01	-0.01	0	-0.01	0	0
Technical-Professional	19	399	0.01	-0.04	-0.05	-0.07	-0.09	-0.21
High School Type:								
Municipal	20	2043	-0.01	-0.06	-0.06	-0.12	-0.05	-0.12
Subsidized	20	3541	0	0	-0.01	0.02	-0.04	-0.1
Private	20	4135	-0.06	-0.05	-0.05	-0.01	-0.01	0.11
Region:								
Center	20	5729	-0.05	-0.08	-0.07	-0.03	-0.06	-0.07
North	19	874	-0.16	-0.15	-0.18	-0.26	-0.32	-0.19
South	20	3116	-0.09	-0.1	-0.1	-0.06	-0.1	-0.11

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 524: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Diseño)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	18	1570	-0.1	-0.09	-0.09	-0.06	-0.01	-0.08
Female	18	2755	0.1	0.09	0.1	0.05	0.04	0.08
SES:								
QA	18	403	0.02	0.03	-	-0.29	0.02	-0.03
QB	17	361	-0.09	-0.09	-	-0.66	-0.06	-0.13
QC	18	408	-0.02	-0.02	-	-0.05	0	0.01
QD	18	315	0.02	-0.03	-	0.16	-0.01	-0.06
QE	18	560	-0.04	-0.04	-	0.17	-0.02	0
Curricular Branch:								
Scientific-Humanistic	18	4055	0	0	0	0.01	0.01	0.02
Technical-Professional	16	270	-0.02	-0.06	-0.04	-0.27	-0.1	-0.24
High School Type:								
Municipal	18	811	0.06	0.01	0.04	-0.1	0.05	-0.06
Subsidized	18	1578	-0.03	-0.06	-0.05	-0.04	-0.08	-0.04
Private	18	1936	-0.02	0	-0.01	0.31	0.05	0.02
Region:								
Center	18	3289	0.03	0.02	0	0.13	0.05	0
North	16	191	-0.29	-0.3	-0.35	0	-0.39	-0.33
South	18	845	0	0	-0.02	-0.1	-0.04	-0.09

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 525: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Educación)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	69	3350	-0.19	-0.18	-0.17	-0.21	-0.1	-0.12
Female	69	5305	0.13	0.12	0.12	0.12	0.06	0.07
SES:								
QA	69	1358	0.03	0.01	-	0.03	-0.03	-0.06
QB	69	2002	-0.04	-0.05	-	0.26	-0.06	-0.12
QC	69	1177	-0.02	-0.02	-	-0.44	0.01	-0.02
QD	69	762	-0.07	-0.05	-	0.12	-0.03	0.03
QE	65	547	-0.03	0.02	-	-0.09	0.07	0.16
Curricular Branch:								
Scientific-Humanistic	69	7051	-0.02	-0.02	-0.02	-0.01	-0.02	0.01
Technical-Professional	69	1604	0.08	0.07	0.06	0.02	0.05	-0.07
High School Type:								
Municipal	69	3746	-0.04	-0.05	-0.05	-0.11	-0.05	-0.09
Subsidized	69	4464	0.02	0.02	0.03	0.01	0.03	0.03
Private	64	445	0.01	0.04	0.06	0.1	0.19	0.34
Region:								
Center	67	3975	0.01	0	-0.02	-0.08	0.06	0.02
North	54	498	-0.19	-0.2	-0.2	-0.32	-0.18	-0.2
South	66	4182	-0.02	-0.01	-0.01	0.01	-0.04	-0.12

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 526: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Educación_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	44	742	-0.3	-0.32	-0.34	-0.28	-0.16	-0.25
Female	58	8816	0.04	0.04	0.04	0.04	0.02	0.03
SES:								
QA	58	1671	0.01	-0.02	-	-0.09	-0.05	-0.1
QB	57	2171	-0.02	-0.01	-	-0.04	-0.04	-0.04
QC	58	1180	-0.03	-0.03	-	-0.03	-0.01	-0.02
QD	58	682	-0.07	-0.07	-	0.05	0	0.04
QE	56	594	0.02	0.03	-	0.21	0.12	0.15
Curricular Branch:								
Scientific-Humanistic	58	7541	-0.02	-0.01	-0.01	-0.04	0	0.01
Technical-Professional	57	2017	0.08	0.03	0.01	0.2	-0.02	-0.07
High School Type:								
Municipal	58	4332	-0.02	-0.03	-0.05	0.05	-0.08	-0.12
Subsidized	58	4273	-0.01	-0.01	0	-0.03	0.01	0.04
Private	51	953	-0.08	0	0.01	0.05	0.16	0.36
Region:								
Center	57	3949	-0.18	-0.17	-0.16	-0.3	-0.13	-0.12
North	39	474	-0.16	-0.16	-0.19	-0.19	-0.15	-0.09
South	57	5135	0.04	0.02	0.05	-0.03	0	0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 527: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Educación_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	36	1688	-0.17	-0.13	-0.17	-0.14	-0.08	-0.08
Female	36	2935	0.09	0.06	0.09	0.08	0.03	0.03
SES:								
QA	36	891	0.01	0.01	-	0.01	-0.01	-0.05
QB	36	1179	0.02	0.01	-	0.01	0	0.03
QC	36	641	-0.05	-0.05	-	-0.02	-0.02	-0.05
QD	35	302	0.06	0.04	-	0.04	0.06	0.09
QE	35	213	0.03	0.07	-	0.05	0.12	0.01
Curricular Branch:								
Scientific-Humanistic	36	4033	-0.01	0	-0.01	0	-0.01	0
Technical-Professional	36	590	0	-0.05	0.02	-0.01	-0.02	-0.02
High School Type:								
Municipal	36	2231	0	-0.02	-0.02	0.01	-0.03	-0.02
Subsidized	36	2267	-0.01	0.01	-0.01	-0.01	0.02	0.01
Private	30	125	-0.16	-0.08	0.35	-0.08	0.11	0.34
Region:								
Center	36	1878	-0.18	-0.16	-0.32	-0.15	-0.01	0.05
North	23	153	-0.25	-0.23	-0.35	-0.26	-0.31	-0.38
South	35	2592	-0.05	-0.04	-0.05	-0.03	-0.06	0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 528: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Educación_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	14	1933	-0.08	-0.08	-0.08	-0.1	-0.05	-0.04
Female	14	1387	0.15	0.15	0.16	0.19	0.11	0.12
SES:								
QA	18	447	0.01	-0.01	-	0	-0.06	-0.05
QB	18	633	-0.05	-0.05	-	0.01	-0.1	-0.1
QC	18	442	0.05	0.06	-	0.04	0.09	0.1
QD	18	319	-0.03	0	-	-0.04	0.02	0.01
QE	17	275	-0.13	-0.07	-	-0.07	-0.04	-0.11
Curricular Branch:								
Scientific-Humanistic	18	2919	-0.02	-0.02	-0.02	-0.02	-0.01	0
Technical-Professional	18	399	0.1	0.08	0.11	0.11	0.06	0.03
High School Type:								
Municipal	18	1210	-0.01	-0.02	-0.01	-0.01	-0.06	-0.1
Subsidized	18	1697	0.04	0.03	0.07	0.04	0.07	0.1
Private	16	411	-0.11	-0.07	-0.46	-0.05	-0.04	0.2
Region:								
Center	17	1499	0.07	0.07	-0.04	0.07	0.1	0.13
North	11	213	0.14	0.19	0.02	0.18	0.14	0.21
South	18	1607	0.04	0.01	-0.1	0.05	0.02	0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 529: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (General)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	768	-0.1	-0.11	-0.14	-0.12	-0.05	-0.09
Female	1	800	0.1	0.11	0.12	0.13	0.05	0.09
SES:								
QA	1	94	-0.08	-0.1	-	-0.07	-0.1	-0.08
QB	1	125	-0.2	-0.19	-	-0.22	-0.21	-0.19
QC	1	132	0.14	0.15	-	0.12	0.14	0.07
QD	1	148	-0.13	-0.13	-	-0.1	-0.09	-0.13
QE	1	229	-0.01	-0.01	-	0	-0.01	-0.02
Curricular Branch:								
Scientific-Humanistic	1	1539	0	0	0.01	-0.01	0	0
Technical-Professional	1	29	-0.1	-0.06	-0.48	0.22	-0.11	-0.09
High School Type:								
Municipal	1	418	0.02	0	-0.04	0.01	0.11	0.06
Subsidized	1	562	-0.06	-0.05	-0.1	-0.04	-0.12	-0.12
Private	1	588	0.04	0.04	0.13	0.03	0.03	0.07
Region:								
Center	1	1314	-0.01	-0.01	0.01	-0.02	0	-0.01
North	1	61	0.13	0.12	-0.06	0.16	0.04	0.1
South	1	193	0.03	0.03	-0.06	0.06	-0.04	0.02

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 530: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Humanidades)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	14	1002	-0.14	-0.13	-0.13	0.04	-0.04	-0.09
Female	14	1648	0.1	0.09	0.09	-0.01	0.05	0.02
SES:								
QA	14	220	-0.06	-0.06	-	-0.16	-0.09	-0.07
QB	14	277	0.01	0	-	-0.79	-0.02	-0.08
QC	14	248	0.05	-0.01	-	-0.32	0.05	-0.06
QD	14	240	-0.19	-0.18	-	0.11	-0.17	-0.19
QE	14	320	-0.05	-0.02	-	0.15	0.16	0.01
Curricular Branch:								
Scientific-Humanistic	14	2499	-0.01	0	0	0.01	0	-0.01
Technical-Professional	14	151	0.08	0.05	0.01	-0.24	0.04	-0.17
High School Type:								
Municipal	14	676	-0.06	-0.08	-0.08	-0.3	-0.06	-0.12
Subsidized	14	1162	0.01	0	0	-0.05	-0.01	-0.05
Private	14	812	0.05	0.09	0.14	-0.16	0.11	0.05
Region:								
Center	14	2123	0.09	0.08	0.11	0.02	0.15	0.06
North	11	71	-0.11	-0.08	-0.16	-1.54	-0.3	-0.21
South	14	456	-0.11	-0.11	-0.12	-0.03	-0.19	-0.16

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 531: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ingeniería_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	125	23572	-0.03	-0.02	-0.03	-0.03	0	0
Female	125	6959	0.11	0.07	0.09	0.11	0.01	0.01
SES:								
QA	124	2796	0.04	0.03	-	0.03	0	-0.03
QB	123	3269	-0.03	-0.07	-	-0.06	-0.1	-0.13
QC	124	2961	0	-0.02	-	-0.01	-0.02	-0.05
QD	124	2455	0.01	0.02	-	0.03	0.02	0.03
QE	123	3279	-0.04	-0.03	-	-0.04	-0.01	0.05
Curricular Branch:								
Scientific-Humanistic	125	28057	-0.02	-0.01	-0.01	-0.01	0	0.02
Technical-Professional	123	2472	0.13	0.03	0.01	0.09	0.03	-0.12
High School Type:								
Municipal	125	8372	0	-0.02	-0.05	-0.03	-0.06	-0.11
Subsidized	125	12555	0	0	0.04	0.01	0	-0.01
Private	121	9602	0.01	0.07	-0.14	0.05	0.15	0.26
Region:								
Center	117	15393	-0.04	-0.03	0.03	-0.04	0.04	0.01
North	97	3262	-0.03	-0.04	-0.04	-0.03	-0.09	-0.12
South	119	11876	-0.02	-0.01	0	0	0.01	-0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 532: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ingeniería_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	64	4795	-0.06	-0.04	-0.03	-0.04	0.01	-0.02
Female	63	2644	0.04	0	0.1	0.05	-0.06	-0.01
SES:								
QA	63	842	0.13	0.09	-	0.08	0.03	0
QB	64	1323	-0.05	-0.06	-	-0.04	-0.08	-0.12
QC	64	1017	-0.04	-0.06	-	-0.06	-0.08	-0.06
QD	62	768	0.13	0.11	-	0.12	0.15	0.06
QE	62	637	-0.03	0.03	-	0.02	0.11	0.12
Curricular Branch:								
Scientific-Humanistic	64	6301	-0.03	-0.01	-0.02	-0.02	-0.01	0.01
Technical-Professional	64	1137	0.12	0.07	0.09	0.08	0.04	-0.03
High School Type:								
Municipal	64	2719	-0.01	-0.03	-0.05	-0.02	-0.02	-0.04
Subsidized	64	3979	0.02	0.02	0.02	0.02	0.02	0.02
Private	59	740	-0.1	-0.02	0.06	-0.04	0.09	0.11
Region:								
Center	59	4392	-0.05	-0.05	0.01	-0.04	-0.01	-0.05
North	43	334	0.15	0.18	0.53	0.15	0.14	0.14
South	62	2713	-0.03	-0.04	-0.13	0.01	-0.03	-0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 533: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Ingeniería_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	13	860	-0.06	-0.05	0.15	-0.08	0.03	0.02
Female	13	861	0.08	0.07	-0.15	0.08	-0.02	-0.02
SES:								
QA	12	136	0.04	0.08	-	0.06	0.02	-0.02
QB	12	156	-0.13	-0.16	-	-0.13	-0.1	-0.15
QC	12	155	-0.05	0	-	-0.03	0.04	-0.02
QD	12	121	-0.11	-0.11	-	-0.13	-0.07	-0.01
QE	12	188	0.22	0.17	-	0.2	0.24	0.26
Curricular Branch:								
Scientific-Humanistic	13	1653	0	0	-0.02	-0.02	0	-0.01
Technical-Professional	9	67	0.24	0.25	0.2	0.22	0.15	0.11
High School Type:								
Municipal	13	514	0.01	-0.02	-0.23	-0.05	-0.05	-0.12
Subsidized	13	812	0.01	0.02	0.19	0.05	0.05	0.02
Private	13	394	0.13	0.17	0.18	0.16	0.18	0.24
Region:								
Center	13	700	0.1	0.13	0.45	0.08	0.15	0.06
North	12	179	0.04	-0.01	-0.47	-0.05	-0.21	-0.13
South	13	842	0.04	0.06	-0.01	0.03	0.01	0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 534: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Mar)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	16	516	-0.06	-0.02	-0.2	-0.06	0.03	-0.01
Female	16	560	0.02	-0.04	0.02	-0.03	-0.11	-0.06
SES:								
QA	16	169	-0.06	-0.08	-	-0.01	-0.13	-0.09
QB	16	184	-0.04	-0.14	-	-0.06	-0.06	-0.16
QC	16	140	0.03	-0.03	-	-0.04	-0.05	-0.02
QD	16	91	-0.14	-0.09	-	-0.27	0.06	-0.16
QE	14	74	0.12	0.22	-	0.12	0.21	0.25
Curricular Branch:								
Scientific-Humanistic	16	978	0	0	0.01	0	0.01	0.03
Technical-Professional	14	98	0.09	0.01	-0.1	0.1	-0.22	-0.27
High School Type:								
Municipal	16	516	-0.07	-0.1	-0.01	-0.03	-0.07	-0.08
Subsidized	16	440	-0.04	-0.07	0.06	-0.08	-0.07	-0.09
Private	16	120	0.27	0.36	-0.37	0.29	0.42	0.69
Region:								
Center	15	215	0.16	0	-0.55	0.07	0.07	-0.17
North	11	129	-0.06	-0.07	-0.84	-0.3	-0.26	-0.36
South	13	732	-0.11	-0.11	0.03	-0.15	-0.06	-0.17

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 535: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Periodismo)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	16	1313	-0.18	-0.18	-0.25	-0.66	-0.06	-0.19
Female	16	2145	0.17	0.17	0.2	0.38	0.08	0.15
SES:								
QA	15	312	-0.05	-0.05	-	0.5	-0.06	-0.11
QB	14	269	0.01	-0.05	-	-0.13	-0.05	-0.18
QC	15	306	-0.12	-0.11	-	-0.21	-0.1	-0.19
QD	16	283	-0.04	-0.01	-	-0.57	0.02	-0.04
QE	15	480	0.15	0.19	-	-0.13	0.2	0.23
Curricular Branch:								
Scientific-Humanistic	16	3306	0	0	0	0.01	0	0.01
Technical-Professional	13	152	0.02	-0.04	-0.03	-	-0.07	-0.24
High School Type:								
Municipal	13	662	-0.13	-0.14	-0.15	-0.16	-0.11	-0.16
Subsidized	16	1336	0	0.02	0.05	-0.07	0	0.05
Private	16	1460	0.05	0.04	0.06	0.29	0.09	0.13
Region:								
Center	16	2188	0	-0.01	0	-0.48	0.05	0.01
North	15	243	-0.14	-0.12	-0.11	-0.11	-0.14	-0.06
South	15	1027	-0.1	-0.09	-0.08	-0.12	-0.08	-0.23

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 536: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Salud_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	31	4705	-0.13	-0.12	-0.17	-0.14	-0.05	-0.08
Female	31	4683	0.12	0.11	0.19	0.13	0.05	0.08
SES:								
QA	29	611	0.07	0.06	-	0.07	0.07	0.06
QB	30	567	-0.09	-0.13	-	-0.14	-0.11	-0.19
QC	28	642	-0.14	-0.16	-	-0.15	-0.13	-0.16
QD	30	671	0.02	0	-	0	0.01	-0.04
QE	30	1073	-0.03	-0.01	-	-0.01	-0.01	0.01
Curricular Branch:								
Scientific-Humanistic	31	9317	0	0	-0.01	0	0	0
Technical-Professional	23	71	-0.14	-0.16	0.27	-0.13	-0.01	-0.4
High School Type:								
Municipal	30	1722	-0.06	-0.07	-0.16	-0.09	0	-0.18
Subsidized	31	3488	-0.01	-0.02	0.03	0	-0.05	-0.08
Private	31	4178	0.02	0.03	-0.01	0.02	0.05	0.13
Region:								
Center	30	4314	-0.2	-0.2	0.09	-0.2	-0.13	-0.11
North	30	1219	0	0	-0.21	0	-0.01	0.01
South	31	3855	-0.08	-0.08	-0.05	-0.07	-0.09	-0.09

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 537: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Salud_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	57	3398	-0.21	-0.2	-0.18	-0.23	-0.13	-0.17
Female	57	10308	0.07	0.06	0.04	0.07	0.04	0.05
SES:								
QA	57	1424	0.03	0.02	-	0.04	-0.02	-0.05
QB	57	2098	-0.03	-0.04	-	-0.04	-0.06	-0.09
QC	57	1820	-0.01	0	-	0	-0.01	-0.05
QD	57	1462	0.02	0.03	-	0.01	0.02	0.04
QE	57	1384	-0.05	-0.05	-	-0.07	0	0.11
Curricular Branch:								
Scientific-Humanistic	57	13133	-0.01	-0.01	-0.01	-0.01	-0.01	0
Technical-Professional	56	573	0.11	0.09	-0.02	0.09	0.13	-0.06
High School Type:								
Municipal	57	4568	0	0.01	-0.07	0	-0.02	-0.05
Subsidized	57	6915	0	0	0	0	0	-0.01
Private	56	2223	-0.01	-0.01	0.02	0	0.06	0.16
Region:								
Center	56	6312	0.03	0.04	-0.07	0.02	0.1	-0.03
North	46	1145	-0.21	-0.2	-0.15	-0.2	-0.23	-0.17
South	55	6249	-0.07	-0.06	-0.01	-0.06	-0.09	-0.08

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 538: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Salud_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	24	1711	-0.21	-0.2	-0.15	-0.25	-0.15	-0.19
Female	24	4378	0.05	0.05	0.04	0.07	0.02	0.04
SES:								
QA	24	631	0.03	0.02	-	0.02	0	-0.03
QB	24	969	-0.09	-0.1	-	-0.1	-0.11	-0.08
QC	24	859	-0.02	-0.01	-	-0.02	-0.01	-0.04
QD	24	618	0.05	0.05	-	0.05	0.09	0.12
QE	24	562	0	0.01	-	0.01	0.04	0.03
Curricular Branch:								
Scientific-Humanistic	24	5805	-0.01	-0.01	-0.02	-0.01	-0.01	0
Technical-Professional	24	282	0.1	0.07	0.27	0.09	0.02	-0.14
High School Type:								
Municipal	24	2064	-0.01	-0.01	0.05	0	-0.02	-0.08
Subsidized	24	3263	0	0	-0.01	0	-0.01	0
Private	24	760	0	0.02	-0.23	-0.01	0.09	0.18
Region:								
Center	24	2986	-0.05	-0.04	0.07	-0.04	0.03	-0.02
North	22	378	-0.24	-0.25	0.11	-0.26	-0.28	-0.14
South	23	2724	-0.03	-0.03	0	-0.01	-0.04	-0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 539: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Técnico_Administración)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	7	504	-0.1	-0.1	-0.06	-0.15	0.02	0.02
Female	7	813	0.07	0.06	0.02	0.13	-0.01	-0.01
SES:								
QA	7	172	0.04	0.05	-	-0.01	-0.02	-0.22
QB	7	245	0.13	0.08	-	-0.1	0.05	-0.13
QC	7	190	-0.09	-0.07	-	0.11	-0.06	0.07
QD	7	142	0.01	0.05	-	-0.12	0.09	0.15
QE	6	112	-0.06	-0.06	-	-0.08	-0.06	-0.02
Curricular Branch:								
Scientific-Humanistic	7	926	-0.12	-0.11	-0.15	-0.04	-0.08	0
Technical-Professional	7	391	0.21	0.18	0.18	0.26	0.16	0.07
High School Type:								
Municipal	7	458	-0.01	-0.01	-0.07	0.02	-0.03	-0.09
Subsidized	7	752	0.03	0.02	0.04	-0.02	0.1	0.08
Private	6	107	-0.14	-0.07	0.05	-0.32	0.1	0.27
Region:								
Center	7	1041	0.01	0.01	0.01	0.04	0.01	0.02
North	6	137	-0.27	-0.31	-0.28	-0.2	-0.38	-0.22
South	7	139	-0.33	-0.26	-0.09	-0.16	-0.31	-0.25

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 540: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Técnico_Agro)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	16	-0.04	-0.05	-0.02	-0.04	-0.04	-
Female	1	8	0.1	0.09	-0.02	0.08	0.1	-
SES:								
QA	1	9	-0.29	-0.29	-	-0.24	-0.27	-
QB	1	8	-0.06	-0.08	-	-0.16	-0.15	-
QC	1	4	0.32	0.29	-	0.33	0.37	-
QD	1	2	0.68	0.63	-	0.68	0.71	-
QE	-	-	-	-	-	-	-	-
Curricular Branch:								
Scientific-Humanistic	1	15	0.12	0.17	0.3	0.15	0.16	-
Technical-Professional	1	9	-0.18	-0.28	-0.49	-0.25	-0.24	-
High School Type:								
Municipal	1	15	-0.11	-0.09	-0.09	-0.08	-0.11	-
Subsidized	1	9	0.21	0.15	0.26	0.14	0.2	-
Private	-	-	-	-	-	-	-	-
Region:								
Center	1	2	-0.53	-0.6	-	-0.63	-0.64	-
North	-	-	-	-	-	-	-	-
South	1	22	0.06	0.05	-0.02	0.06	0.07	-

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 541: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Técnico_Ciencias)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	4	199	-0.11	-0.07	-0.04	-0.12	-0.06	-0.05
Female	4	280	0.08	0.05	0.06	0.1	0.04	0.02
SES:								
QA	4	57	-0.07	-0.09	-	-0.01	-0.13	-0.15
QB	4	109	0.09	0.16	-	0.1	0.11	0.01
QC	4	55	-0.11	-0.15	-	-0.08	-0.18	-0.12
QD	4	44	0.25	0.18	-	0.17	0.25	0.35
QE	4	28	-0.18	-0.07	-	-0.22	0.01	0.11
Curricular Branch:								
Scientific-Humanistic	4	429	-0.02	0	0.07	-0.01	-0.01	0.01
Technical-Professional	4	50	0.04	-0.01	-0.48	0.01	0.04	-0.17
High School Type:								
Municipal	4	282	-0.03	-0.03	-0.11	-0.04	-0.05	-0.1
Subsidized	4	185	0.05	0.05	0.44	0.07	0.06	0.12
Private	3	12	-0.02	0.04	0.02	-0.03	0.32	0.34
Region:								
Center	4	156	0.35	0.36	-	0.37	0.47	0.3
North	2	7	0.02	-0.09	-	-0.1	-0.06	-0.33
South	4	316	0.15	0.11	0.02	0.12	0.11	0.27

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 542: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Técnico_Diseño)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	3	155	0.03	0.05	0.11	0.14	0.04	0.06
Female	3	122	0.04	0.04	-0.14	-0.14	0.09	0.12
SES:								
QA	3	30	-0.1	-0.14	-	0.05	-0.07	-0.14
QB	3	64	0.2	0.15	-	0.01	0.13	0.08
QC	3	39	-0.09	-0.03	-	0.26	0.03	0.07
QD	3	34	-0.15	-0.1	-	0.03	-0.16	-0.11
QE	3	25	0.16	0.18	-	-0.49	0.25	0.28
Curricular Branch:								
Scientific-Humanistic	3	214	-0.1	-0.06	-0.19	-0.11	-0.04	-0.01
Technical-Professional	3	63	0.28	0.17	0.44	0.39	0.15	0.02
High School Type:								
Municipal	3	100	0.06	0.02	0	0.19	0	-0.03
Subsidized	3	167	0.01	0.01	0.05	-0.04	0.03	0.02
Private	3	10	-0.32	-0.35	-0.71	-1.3	-0.2	0.07
Region:								
Center	3	261	-0.02	-0.02	-0.05	-0.03	-0.01	-0.01
North	2	1	1.59	1.71	1.58	1.23	1.56	0.98
South	3	15	0.13	0.11	0.2	-	0.14	0.25

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 543: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Técnico_Educación)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	44	-0.14	-0.14	-0.07	-0.08	-0.13	-0.15
Female	1	29	0.29	0.28	0.36	0.16	0.25	0.23
SES:								
QA	1	9	0.27	0.18	-	0.26	0.12	0.18
QB	1	20	-0.14	-0.14	-	-0.08	-0.11	-0.11
QC	1	10	-0.04	-0.1	-	-0.02	-0.04	-0.03
QD	1	7	-0.26	-0.26	-	-0.56	-0.23	-0.47
QE	1	7	-0.08	0.23	-	-0.06	0.05	-0.11
Curricular Branch:								
Scientific-Humanistic	1	60	0.04	0.04	-0.07	0	0.03	0.03
Technical-Professional	1	13	-0.19	-0.22	0.29	-0.04	-0.18	-0.15
High School Type:								
Municipal	1	23	-0.05	-0.04	-0.19	0.1	0	-0.06
Subsidized	1	48	0.02	0.03	0.11	-0.05	-0.02	0.03
Private	1	2	0.06	-0.18	0.84	-0.04	0.01	-0.15
Region:								
Center	1	56	-0.07	-0.08	-0.02	-0.04	-0.08	-0.03
North	1	7	0.27	0.41	0.84	0.28	0.39	0.1
South	1	10	0.27	0.28	-0.03	0.03	0.23	0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 544: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Técnico_Idioma)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	6	215	-0.31	-0.2	-0.16	-0.04	-0.09	0.08
Female	6	445	0.14	0.13	0.13	0.03	0.07	-0.05
SES:								
QA	6	73	0.1	-0.03	-	-0.06	-0.08	-0.15
QB	6	116	-0.03	0	-	0.1	-0.02	0.03
QC	6	90	-0.12	-0.07	-	-0.29	0.02	-0.09
QD	5	59	-0.1	-0.16	-	-0.27	-0.08	0.1
QE	4	47	-0.08	0.05	-	-0.15	-0.06	-0.06
Curricular Branch:								
Scientific-Humanistic	6	577	-0.02	-0.03	-0.02	0.02	-0.05	0.01
Technical-Professional	6	83	0.12	0.14	0.12	-0.22	0.2	-0.08
High School Type:								
Municipal	6	258	0.08	0.08	0.05	-0.09	0.05	-0.1
Subsidized	6	338	-0.1	-0.11	-0.12	0.01	-0.08	0.01
Private	4	64	0.26	0.24	0.28	0.39	0.36	0.33
Region:								
Center	6	276	0.19	0.15	0.17	0.01	0.13	0.1
North	6	24	-0.06	0.05	-0.03	0.29	0.03	-0.32
South	6	360	-0.12	-0.17	0.12	-0.1	-0.11	0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 545: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Técnico_Ingeniería)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	61	5697	-0.04	-0.03	0	-0.05	-0.01	-0.03
Female	58	1600	-0.03	-0.08	-0.07	-0.03	-0.1	-0.04
SES:								
QA	61	973	-0.02	-0.01	-	0.03	-0.05	-0.03
QB	60	1537	0.04	0.01	-	0.01	0.01	-0.02
QC	61	1128	0.01	0.02	-	0.03	0.02	-0.03
QD	61	660	0.02	-0.01	-	-0.06	0	0.06
QE	59	498	-0.17	-0.17	-	-0.24	-0.09	0.01
Curricular Branch:								
Scientific-Humanistic	61	5175	-0.08	-0.06	-0.08	-0.06	-0.03	0
Technical-Professional	60	2121	0.16	0.12	0.1	0.1	0.07	0.03
High School Type:								
Municipal	61	3027	0	-0.01	0.02	-0.01	-0.04	-0.07
Subsidized	61	3867	0.01	0.03	0.04	0.03	0.04	0.05
Private	54	402	-0.08	0	-0.52	-0.07	0.14	0.25
Region:								
Center	57	4741	-0.01	-0.01	-0.05	-0.11	0.05	0.05
North	46	315	-0.1	-0.1	-0.13	-0.04	-0.16	-0.1
South	59	2241	0	0.01	0.05	-0.01	-0.03	0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Table 546: Average Over Prediction (-) and Under Prediction (+) of SYGPA by Predictor Measure and Subgroups (Veterinaria)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	5	910	-0.06	-0.06	-0.12	-0.07	0.02	-0.01
Female	5	1329	0.05	0.04	0.09	0.05	-0.01	0.01
SES:								
QA	5	196	-0.02	-0.02	-	-0.03	-0.04	-0.02
QB	5	252	-0.09	-0.09	-	-0.08	-0.1	-0.13
QC	5	259	-0.01	0.01	-	-0.01	-0.01	-0.05
QD	5	222	-0.08	-0.12	-	-0.12	-0.08	-0.14
QE	5	230	0.06	0.06	-	0.07	0.1	0.09
Curricular Branch:								
Scientific-Humanistic	5	2130	0	0	-0.02	0	0	0
Technical-Professional	5	109	0.09	0.05	0.2	0.09	0.04	-0.06
High School Type:								
Municipal	5	644	-0.04	-0.05	0.05	-0.04	-0.05	-0.12
Subsidized	5	1042	-0.01	-0.02	-0.06	-0.02	-0.04	-0.04
Private	5	553	0.04	0.05	-0.06	0.04	0.14	0.2
Region:								
Center	5	1038	-0.08	-0.08	0.04	-0.09	-0.02	-0.1
North	5	138	0	-0.01	0.12	-0.01	-0.05	-0.01
South	5	1063	-0.07	-0.07	0.1	-0.07	-0.07	-0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed SYGPA. Prediction equations are estimated within careers.)

Appendix X. Prediction Bias by the Type of Career - University Completion

Table 547: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Administración)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	35	3272	-0.02	-0.02	-0.02	0	0.01	-0.02
Female	35	2602	0.03	0.03	0.02	0.01	-0.03	-0.01
SES:								
QA	35	851	-0.01	0.01	-0.02	0.12	-0.04	-0.08
QB	35	837	0.05	0.05	0.01	-0.01	0.03	-0.01
QC	35	618	-0.03	-0.02	-0.02	0.07	-0.03	0
QD	35	433	-0.14	-0.13	-0.11	-0.07	-0.12	-0.11
QE	35	648	0.22	0.21	0.2	0.11	0.21	0.18
Curricular Branch:								
Scientific-Humanistic	35	5030	-0.03	-0.02	-0.03	-0.02	-0.02	-0.01
Technical-Professional	34	843	0.19	0.19	0.12	0.16	0.13	0.05
High School Type:								
Municipal	35	1465	0.1	0.12	0.11	0.16	0.09	0.08
Subsidized	35	1806	-0.03	-0.02	-0.04	0.01	-0.03	-0.05
Private	34	2602	-0.08	-0.06	-0.04	-0.14	-0.05	-0.04
Region:								
Center	35	3392	-0.1	-0.09	-0.1	-0.13	-0.06	0
North	22	559	-0.09	-0.09	-0.11	0.01	-0.12	-0.05
South	35	1922	0.01	-0.02	0.02	-0.01	-0.01	0

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 548: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Administración_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	28	1137	-0.01	-0.01	-0.02	-0.09	-0.01	0.03
Female	28	1290	0.01	-0.01	0.02	0.09	0	-0.01
SES:								
QA	28	586	0.08	0.11	0.09	0.2	0.1	-0.01
QB	28	806	-0.03	-0.09	-0.08	0.1	-0.06	0.01
QC	27	370	-0.05	0	0.04	-0.16	0.03	-0.01
QD	22	178	-0.14	-0.11	0	-0.22	-0.11	-0.07
QE	22	115	0.15	0.14	0.11	0.01	0.18	0.27
Curricular Branch:								
Scientific-Humanistic	28	1330	-0.09	-0.05	-0.07	-0.03	-0.03	-0.01
Technical-Professional	28	1096	0.06	0.03	0.03	0.06	0.02	0.03
High School Type:								
Municipal	28	1269	-0.02	-0.02	-0.01	-0.03	-0.01	0
Subsidized	28	1014	0.02	0.02	0.01	0	0.02	0.04
Private	16	143	-0.34	-0.31	-0.33	-0.33	-0.24	-0.2
Region:								
Center	24	1250	0.12	0.11	0.08	0	0.16	0.14
North	11	153	-0.09	-0.2	-0.19	-0.01	-0.15	-0.04
South	26	1023	0.09	0.12	0.15	-0.11	0.13	0.19

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 549: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Administración_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	7	202	-0.06	-0.08	-0.11	-0.19	-0.02	0.01
Female	7	413	0.06	0.05	0.08	0.1	0.07	0.02
SES:								
QA	7	88	-0.03	0.01	0.05	-0.16	-0.03	0.06
QB	7	124	-0.01	-0.03	-0.03	0.27	0.29	-0.05
QC	7	100	-0.02	-0.04	-0.05	-0.19	-0.05	0.02
QD	7	52	0.01	-0.02	0.05	-0.02	-0.02	-0.01
QE	7	72	-0.19	-0.25	-0.2	-0.24	-0.1	-0.31
Curricular Branch:								
Scientific-Humanistic	7	550	0.02	0.01	0.02	0.06	0.03	0.01
Technical-Professional	7	64	-0.01	0.12	0.1	-1.39	0.22	-0.04
High School Type:								
Municipal	7	155	-0.21	-0.23	-0.14	-0.3	-0.15	-0.22
Subsidized	7	308	0.08	0.05	0.05	0.12	0.11	0.05
Private	7	151	0.15	0.26	0.19	0.33	0.16	0.25
Region:								
Center	7	446	-0.02	-0.01	-0.01	0.07	0.02	0
North	6	19	0	0.05	0.1	-1.43	-0.17	0.12
South	7	149	-0.14	-0.17	-0.11	0.17	-0.16	-0.11

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 550: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Agro)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	15	1362	0.04	0.05	0.01	0.05	0.06	0.08
Female	15	1043	-0.08	-0.1	-0.06	-0.1	-0.09	-0.15
SES:								
QA	15	364	0.06	0.06	0.11	0.07	0.11	0.16
QB	15	434	-0.02	-0.01	0	-0.03	-0.03	-0.05
QC	15	334	-0.13	-0.08	-0.12	-0.1	-0.1	-0.14
QD	15	235	-0.03	0	0.03	0	-0.01	-0.01
QE	14	275	0.05	-0.01	-0.08	0	0.02	-0.02
Curricular Branch:								
Scientific-Humanistic	15	2238	0.03	0.03	0.02	0.02	0.01	0
Technical-Professional	15	166	-0.13	-0.1	-0.13	-0.07	-0.02	-0.02
High School Type:								
Municipal	15	872	0.02	0.02	0.03	0.02	0.02	0
Subsidized	15	928	-0.04	-0.03	-0.05	-0.03	-0.02	-0.04
Private	15	604	-0.04	-0.12	0.02	-0.06	-0.07	-0.01
Region:								
Center	15	1277	-0.02	-0.02	-0.14	0	0.08	-0.04
North	12	80	-0.05	-0.07	-0.02	0	-0.07	0.09
South	15	1047	0.07	0.08	0.04	0.06	0.02	0.09

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 551: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Agro_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	4	125	-0.06	-0.04	-0.04	-0.03	-0.01	-0.04
Female	4	133	0.07	0.04	0.06	0.01	0	0.04
SES:								
QA	4	43	-0.13	-0.19	0.07	0.14	-0.15	-0.27
QB	4	72	0.4	0.33	0.03	0.33	0.35	0.48
QC	4	45	-0.2	-0.17	-0.03	-0.3	-0.14	-0.1
QD	4	33	-0.19	-0.11	-0.17	-0.22	-0.21	-0.18
QE	4	15	-0.42	-0.21	-0.29	-0.28	-0.26	-0.06
Curricular Branch:								
Scientific-Humanistic	4	216	-0.01	0	0	0.01	0.01	0.03
Technical-Professional	4	42	0.05	0	-0.01	-0.08	-0.03	-0.07
High School Type:								
Municipal	4	92	-0.01	-0.01	0.02	-0.02	-0.09	-0.03
Subsidized	4	135	-0.06	-0.1	-0.12	-0.1	-0.04	-0.05
Private	4	31	-0.29	-0.02	-0.02	-0.27	-0.18	-0.22
Region:								
Center	4	175	-0.29	-0.28	-0.27	-0.22	-0.08	-0.02
North	3	4	-0.46	-0.24	0.11	-0.47	-0.33	-0.42
South	4	79	0.33	0.32	0.47	1.32	0.32	0.5

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 552: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Arquitectura)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	18	1680	-0.01	-0.01	-0.02	-0.01	0	-0.01
Female	18	1585	0.01	0.01	0.03	0	-0.01	-0.01
SES:								
QA	18	351	-0.01	0	0.01	0.05	-0.03	-0.01
QB	18	352	-0.05	-0.04	-0.04	-0.01	-0.03	-0.06
QC	18	292	-0.05	-0.06	-0.06	-0.07	-0.04	0.01
QD	18	272	0.15	0.12	0.18	0.06	0.14	0.04
QE	18	382	-0.05	-0.04	-0.03	0.08	-0.07	-0.08
Curricular Branch:								
Scientific-Humanistic	18	3068	0	0	0.01	0	0	-0.01
Technical-Professional	17	196	0.09	0.13	0.06	0.17	0.09	0.03
High School Type:								
Municipal	18	683	0.07	0.08	0.03	0.06	0.04	0.05
Subsidized	18	1220	-0.07	-0.08	-0.06	-0.08	-0.06	-0.09
Private	18	1361	0.06	0.06	0.09	-0.07	0.07	0.01
Region:								
Center	18	1806	0.03	0.06	0.03	0.07	0.05	0.11
North	16	384	-0.01	0.01	0.03	-0.24	0.03	-0.05
South	18	1074	-0.1	-0.12	-0.1	-0.07	-0.11	-0.14

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 553: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Arte_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	12	213	-0.17	-0.16	-0.15	-0.21	-0.09	-0.06
Female	12	766	0.07	0.07	0.07	0.14	0.04	0.02
SES:								
QA	11	106	-0.09	-0.15	-0.12	-0.4	-0.07	-0.16
QB	11	83	-0.27	-0.28	-0.25	-0.31	-0.26	-0.18
QC	12	61	-0.08	-0.1	-0.11	-0.56	-0.12	-0.02
QD	12	72	0.12	0.03	0.07	0.64	0.09	-0.01
QE	12	149	0.12	0.08	0.1	-0.48	0.14	0.19
Curricular Branch:								
Scientific-Humanistic	12	962	0.02	0.03	0.03	0.02	0.01	0.01
Technical-Professional	10	17	0	-0.02	-0.08	-0.69	0.01	0.21
High School Type:								
Municipal	11	132	-0.1	-0.15	-0.13	0.07	-0.13	-0.28
Subsidized	12	283	-0.03	-0.05	-0.06	-0.2	-0.07	-0.06
Private	12	564	0.07	0.09	0.07	0.74	0.06	0.09
Region:								
Center	11	703	-0.06	0.05	-0.06	-0.33	0	0.01
North	10	40	0.08	0.05	0.04	-0.31	0.04	0.19
South	12	236	-0.14	-0.12	-0.15	0.17	-0.19	-0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 554: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Arte_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	10	276	-0.05	-0.02	-0.04	-0.02	0.01	-0.05
Female	10	333	0.01	-0.01	0.02	-0.11	-0.04	0
SES:								
QA	9	48	0.12	0.13	0.04	0.32	0.06	-0.24
QB	9	56	-0.11	-0.04	-0.1	0.09	-0.06	0.16
QC	10	45	0.04	-0.13	-0.03	0.49	0	-0.27
QD	10	54	0.1	0.04	0.06	-0.15	0.09	-0.08
QE	9	95	-0.12	-0.09	-0.15	-0.28	-0.07	-0.01
Curricular Branch:								
Scientific-Humanistic	10	599	0.01	0.01	0.01	-0.02	0.01	0.02
Technical-Professional	8	10	-0.16	-0.23	-0.18	-	-0.09	-0.64
High School Type:								
Municipal	10	96	0.13	0.15	0.09	-0.39	0.16	0.09
Subsidized	10	200	0	0	0.05	0.29	-0.03	-0.09
Private	10	313	0.11	0.08	0.13	-0.07	0.16	0.19
Region:								
Center	10	501	0.05	0	0.02	-0.09	0.06	0
North	9	24	0.17	0.2	0.15	0.15	0.19	0.22
South	10	84	-0.06	0	-0.02	0.42	-0.04	-0.09

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 555: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ciencias)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	15	747	-0.09	-0.06	-0.12	-0.1	-0.07	-0.07
Female	15	906	0.06	0.06	0.14	0.07	0.05	0.02
SES:								
QA	15	148	-0.04	-0.07	0.31	-0.07	-0.07	-0.1
QB	15	273	0.08	0.07	0.17	0.11	0.11	0.09
QC	15	202	-0.08	-0.08	0.04	-0.02	-0.09	-0.01
QD	15	162	0.12	0.11	0.04	0.11	0.09	0.09
QE	13	155	-0.04	-0.06	-0.39	-0.01	-0.02	0.06
Curricular Branch:								
Scientific-Humanistic	15	1575	0.01	0	0.02	0.01	0.01	0
Technical-Professional	13	78	-0.07	0.08	-0.15	-0.07	-0.08	-0.03
High School Type:								
Municipal	15	472	-0.05	-0.05	-0.02	-0.05	-0.05	-0.04
Subsidized	15	729	0.13	0.11	0.03	0.13	0.11	0.09
Private	15	452	-0.06	0.01	0.02	0	-0.04	0.01
Region:								
Center	14	1226	0.39	0.4	-0.08	0.39	0.37	0.16
North	11	92	0.08	0.13	0.15	0.13	0.04	0.22
South	14	335	-0.09	-0.02	0.23	-0.1	-0.1	-0.1

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 556: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ciencias_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	25	1208	-0.02	-0.01	0.07	-0.02	0.01	-0.01
Female	25	1567	0.02	0.01	-0.09	0.02	0	0.01
SES:								
QA	25	347	0.14	0.12	0.05	0.13	0.15	0.11
QB	25	604	-0.01	0	-0.09	0	-0.01	-0.02
QC	25	432	-0.02	-0.01	0.28	0	-0.01	-0.04
QD	25	345	-0.02	-0.02	-0.29	-0.03	-0.02	-0.03
QE	25	244	-0.08	-0.06	0.32	-0.07	-0.06	0.03
Curricular Branch:								
Scientific-Humanistic	25	2665	0	0.01	0	0.01	0.01	0.01
Technical-Professional	24	109	0.07	-0.07	-0.19	0.04	0.02	-0.08
High School Type:								
Municipal	25	995	-0.04	-0.03	-0.11	-0.04	-0.02	-0.05
Subsidized	25	1345	0.03	0	0	0.03	0.01	0
Private	24	434	0.01	0.02	0.34	0.01	0.02	0.01
Region:								
Center	24	1463	0.01	-0.02	0.04	0.01	0	-0.12
North	23	364	0	-0.04	-0.22	-0.03	-0.07	0.01
South	24	947	-0.07	-0.05	0.04	-0.06	-0.08	-0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 557: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ciencias_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	13	559	-0.04	-0.03	-0.08	-0.04	-0.02	-0.05
Female	13	468	0.03	0	0.06	0.02	0	0.08
SES:								
QA	13	172	0.2	0.18	0.04	0.2	0.22	0.34
QB	13	228	0	-0.02	0.14	-0.02	-0.02	0.04
QC	13	152	0.02	0.06	-0.16	0.04	0.01	-0.24
QD	13	93	0.15	0.12	0.02	0.1	0.15	0.08
QE	12	83	0.26	0.27	0.57	0.27	0.25	0.19
Curricular Branch:								
Scientific-Humanistic	13	936	-0.01	-0.01	0.02	-0.02	0	0
Technical-Professional	12	91	0.32	0.26	-0.14	0.31	0.29	0.12
High School Type:								
Municipal	13	375	0.02	-0.01	-0.1	-0.01	0.03	-0.08
Subsidized	13	487	-0.04	-0.03	0.14	-0.05	-0.04	0.01
Private	12	165	-0.08	-0.05	-0.2	0.02	-0.07	-0.1
Region:								
Center	12	493	0.39	0.4	-0.09	0.36	0.39	0.08
North	12	184	-0.13	-0.1	0.2	-0.12	0.02	-0.19
South	13	350	0.1	0.05	0.05	0.08	0.08	0.19

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 558: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ciencias_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	2	146	0.02	0.05	0.07	0.04	0.06	0.14
Female	2	80	-0.06	-0.09	-0.12	-0.08	-0.12	-0.19
SES:								
QA	2	26	0.17	0.13	-0.19	0.18	0.21	0.26
QB	2	33	-0.14	-0.18	-0.21	-0.13	-0.12	-0.12
QC	2	38	0.01	0.01	-0.17	-0.01	0.02	0.08
QD	2	19	-0.2	-0.15	-0.16	-0.2	-0.22	-0.24
QE	2	26	-0.25	-0.19	-	-0.23	-0.25	-0.19
Curricular Branch:								
Scientific-Humanistic	2	212	0.01	0.01	0.07	0.01	0.01	0.01
Technical-Professional	2	14	-0.08	-0.12	-0.3	-0.12	-0.11	-0.14
High School Type:								
Municipal	2	75	-0.04	-0.07	-0.2	-0.05	-0.04	-0.04
Subsidized	2	97	-0.07	-0.05	0.66	-0.07	-0.08	0.07
Private	2	54	0.25	0.29	-0.33	0.28	0.27	0
Region:								
Center	2	18	-0.15	-0.24	-0.24	-0.29	-0.26	-0.23
North	2	154	-0.09	0.02	-	-0.04	-0.06	0
South	2	54	-0.05	-0.14	0.03	-0.05	-0.05	0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 559: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ciencias_Sociales_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	32	1492	-0.08	-0.07	-0.07	-0.11	-0.03	-0.05
Female	32	2593	0.06	0.05	0.05	0.08	0.02	0.03
SES:								
QA	32	486	0.11	0.1	0.12	0.07	0.1	0.07
QB	32	549	-0.06	-0.05	-0.07	-0.1	-0.09	-0.13
QC	32	442	-0.04	-0.03	-0.03	-0.2	-0.04	-0.04
QD	32	386	0.02	0	0.01	-0.01	0.03	0.03
QE	32	475	-0.12	-0.1	-0.08	-0.08	-0.08	-0.13
Curricular Branch:								
Scientific-Humanistic	32	3873	0.01	0.01	0.02	0.01	0.02	0.01
Technical-Professional	28	212	-0.09	-0.12	-0.14	-0.11	-0.16	-0.21
High School Type:								
Municipal	31	970	-0.06	-0.05	-0.06	-0.11	-0.05	-0.06
Subsidized	32	1449	-0.03	-0.04	-0.03	-0.06	-0.05	-0.05
Private	31	1666	0.13	0.12	0.13	-0.02	0.15	0.02
Region:								
Center	32	2143	-0.04	-0.04	-0.06	-0.2	-0.02	-0.1
North	26	463	-0.05	-0.06	-0.03	-0.18	-0.11	-0.1
South	32	1479	-0.05	-0.05	-0.05	0.03	-0.04	-0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 560: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ciencias_Sociales_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	21	1204	-0.04	-0.01	-0.05	-0.19	0.02	-0.03
Female	21	1263	0.04	0.04	0.04	0.18	0.01	0.05
SES:								
QA	21	313	0.1	0.07	0.07	-0.24	0.06	0.18
QB	21	464	-0.05	-0.02	-0.05	-0.37	-0.03	-0.07
QC	21	347	-0.01	-0.03	-0.01	0.15	0.02	0
QD	21	257	0.1	0.08	0.12	0.59	0.09	0.1
QE	21	279	-0.01	0	0.01	-0.21	0.02	0.06
Curricular Branch:								
Scientific-Humanistic	21	2318	-0.01	-0.01	-0.01	0.02	0	0
Technical-Professional	21	149	0.03	-0.02	-0.01	-0.15	0.04	-0.01
High School Type:								
Municipal	21	726	-0.05	-0.07	-0.05	-0.13	-0.05	-0.09
Subsidized	21	952	0.04	0.06	0.04	0.13	0.06	0.07
Private	20	789	0.01	0.03	0.02	0.46	0.03	0.14
Region:								
Center	20	1840	-0.08	-0.09	-0.09	-0.06	-0.07	-0.07
North	17	121	-0.08	-0.04	-0.1	-0.49	-0.13	-0.06
South	21	506	-0.02	-0.02	-0.02	0.08	-0.02	0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 561: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ciencias_Sociales_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	17	258	-0.01	0.03	0.01	-0.38	0.04	0.07
Female	17	1351	0.01	0	0	0.1	0	0
SES:								
QA	17	337	-0.02	0	-0.05	-0.01	-0.04	-0.13
QB	17	449	0.09	0.08	0.11	0.31	0.07	0.14
QC	17	283	0.07	0.07	0.06	-0.5	0.08	0.12
QD	17	157	-0.03	-0.02	0.02	0.16	-0.01	0.05
QE	16	108	-0.11	-0.11	-0.09	0.27	-0.09	-0.03
Curricular Branch:								
Scientific-Humanistic	17	1375	-0.01	-0.01	-0.01	-0.08	0	0.01
Technical-Professional	17	234	0.05	0.08	0.05	0.49	0.05	0.05
High School Type:								
Municipal	17	709	0.03	0.04	0.03	-0.02	0.03	0.08
Subsidized	17	738	0.01	0.01	0.01	-0.02	0.02	0
Private	15	162	-0.05	-0.03	0	0.05	-0.02	-0.18
Region:								
Center	16	746	-0.1	-0.07	-0.1	0.05	-0.08	0.02
North	14	169	-0.11	-0.06	-0.04	0.4	-0.15	0.03
South	17	694	0.06	0.01	-0.04	-0.24	0.03	-0.02

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 562: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Comunicaciones)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	2	92	-0.12	-0.2	-0.12	0	-0.12	-0.19
Female	2	58	0.16	0.2	0.16	0.15	0.15	0.21
SES:								
QA	2	16	-0.01	-0.07	-0.08	-1.17	-0.06	-0.01
QB	2	24	0.46	0.45	0.45	-0.07	0.43	0.49
QC	2	23	-0.89	-0.96	-1.08	-0.75	-1.06	-0.86
QD	2	18	-0.13	-0.16	-0.2	0.33	-0.19	-0.17
QE	2	18	0.11	0.13	0.3	0.17	0.15	0.11
Curricular Branch:								
Scientific-Humanistic	2	140	0	0.01	-0.02	0.03	0	0.01
Technical-Professional	2	10	-0.66	-0.71	0.22	-0.53	-0.72	-0.49
High School Type:								
Municipal	2	35	0.14	0.13	0.1	0.13	0.24	0.08
Subsidized	2	78	0.05	-0.04	0.05	0.01	-0.05	0.03
Private	2	37	-0.05	0.06	-0.06	0.17	-0.01	0.01
Region:								
Center	2	105	0	0.02	0	-0.08	0.04	-0.01
North	2	17	0.17	0.05	0.24	0.06	-0.03	0.14
South	2	28	-0.03	-0.04	-0.12	-0.03	-0.11	0

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 563: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Construcción)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	15	1975	-0.01	-0.01	-0.03	-0.01	0	-0.01
Female	15	634	0.03	0.01	0.12	0.02	-0.01	0
SES:								
QA	15	423	-0.07	-0.11	-0.07	-0.1	-0.08	-0.13
QB	15	626	0.02	0.02	0.07	0.03	0.02	0
QC	15	445	-0.06	-0.06	-0.04	-0.07	-0.08	-0.09
QD	15	275	0.01	0.06	0.14	0.06	0.07	0.11
QE	15	211	0.05	0.07	-0.14	0.04	0.04	0.15
Curricular Branch:								
Scientific-Humanistic	15	2357	-0.01	0	-0.01	0	0	0
Technical-Professional	15	252	0.02	0	0.06	-0.01	0	0.03
High School Type:								
Municipal	15	895	0.06	0.06	0.23	0.06	0.04	0.03
Subsidized	15	1383	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01
Private	15	331	-0.08	-0.09	-0.14	-0.06	0.02	-0.01
Region:								
Center	15	975	-0.04	-0.03	-0.09	-0.04	-0.04	0.07
North	11	837	-0.15	-0.03	0.18	-0.08	-0.13	-0.18
South	15	797	0.05	0.07	-0.02	0	0.05	-0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 564: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Derecho)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	21	2904	-0.01	-0.01	-0.01	-0.03	0.01	-0.01
Female	21	2783	0.02	0.02	0.02	-0.01	0	0.01
SES:								
QA	21	671	0.01	0.01	0.01	0.05	0	0.01
QB	21	719	-0.08	-0.07	-0.07	-0.06	-0.07	-0.1
QC	21	633	-0.06	-0.05	-0.06	-0.12	-0.05	-0.06
QD	21	526	-0.03	-0.04	-0.02	-0.13	-0.05	-0.04
QE	21	667	0.19	0.19	0.19	0.42	0.22	0.21
Curricular Branch:								
Scientific-Humanistic	21	5399	0	0	0	0.02	0	0
Technical-Professional	19	287	-0.11	-0.11	-0.11	-0.43	-0.11	-0.14
High School Type:								
Municipal	21	1286	-0.03	-0.04	-0.04	-0.02	-0.03	-0.03
Subsidized	21	2171	-0.05	-0.06	-0.06	-0.1	-0.05	-0.06
Private	21	2229	0.02	0.04	0.02	0.33	0.02	0.04
Region:								
Center	21	2929	-0.06	-0.07	-0.08	-0.08	-0.06	-0.07
North	20	932	-0.09	-0.07	-0.1	0.01	-0.07	-0.05
South	21	1825	0.04	0.04	0.04	-0.07	0.03	0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 565: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Diseño)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	20	911	-0.04	-0.02	-0.02	-0.01	0.01	0.02
Female	20	1545	0.01	0	-0.01	0.07	-0.02	-0.02
SES:								
QA	20	270	-0.01	-0.03	-0.04	-0.16	-0.08	-0.06
QB	20	273	0.01	-0.03	-0.05	0	-0.03	-0.05
QC	20	278	0	0.01	-0.03	0	0.02	0.05
QD	20	205	-0.04	-0.01	-0.1	0	-0.03	0.1
QE	20	357	0.02	0	0.03	0.03	-0.02	-0.04
Curricular Branch:								
Scientific-Humanistic	20	2315	0	0.01	0	0.06	0	0.02
Technical-Professional	18	140	0.03	0.03	0.07	-0.43	0.03	-0.15
High School Type:								
Municipal	20	438	-0.01	-0.01	-0.07	-0.24	0	0.01
Subsidized	20	864	0.02	0.02	0.01	0.02	0	-0.01
Private	19	1153	-0.02	-0.02	0	0.05	-0.04	0.02
Region:								
Center	20	1920	-0.03	-0.03	-0.03	-0.08	-0.03	-0.02
North	19	141	-0.03	-0.03	-0.12	0.2	-0.08	-0.04
South	20	394	-0.01	0	-0.03	0.02	-0.02	-0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 566: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Educación)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	67	1761	-0.05	-0.05	-0.05	-0.2	-0.01	-0.03
Female	67	2805	0.04	0.04	0.04	0.16	0.02	0.04
SES:								
QA	67	949	0.04	0.05	0.02	0.08	0.01	0.01
QB	67	1435	0.02	0.01	0.02	0.05	0.02	-0.01
QC	67	750	-0.04	-0.04	-0.03	-0.71	-0.04	-0.01
QD	67	413	0	-0.01	-0.01	0.82	0.02	0.01
QE	65	262	0.07	0.07	0.11	-0.74	0.11	0.17
Curricular Branch:								
Scientific-Humanistic	67	3776	-0.01	-0.01	-0.01	0.05	0	0.01
Technical-Professional	67	789	0.01	0.02	0.02	-0.44	0	-0.05
High School Type:								
Municipal	67	2011	0	0.01	0	-0.14	0	-0.02
Subsidized	67	2298	0	0.01	0.01	-0.01	0.01	0.01
Private	63	256	0.02	0.05	0	0.04	0.08	0.2
Region:								
Center	66	2009	-0.04	-0.05	-0.04	0.15	-0.02	-0.05
North	43	593	-0.09	-0.11	-0.09	-1.65	-0.14	0.04
South	65	1963	0	0.04	0.02	0.04	-0.02	0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 567: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Educación_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	43	313	-0.32	-0.28	-0.23	-0.35	-0.25	-0.37
Female	56	4028	0.02	0.02	0.01	0.02	0.01	0.01
SES:								
QA	55	1004	0.05	0.03	0.09	-0.14	0.03	0.09
QB	56	1355	-0.06	-0.07	-0.04	-0.06	-0.08	-0.07
QC	56	640	0.01	0	-0.02	0	0.03	0.03
QD	55	350	0.04	0.02	0.03	0.24	0.05	-0.01
QE	52	286	0	-0.04	-0.01	-0.11	0.01	-0.01
Curricular Branch:								
Scientific-Humanistic	56	3585	0	0.01	0	0	0.01	0.01
Technical-Professional	54	756	0.01	-0.02	-0.02	-0.08	-0.01	-0.07
High School Type:								
Municipal	56	2008	-0.06	-0.05	-0.06	0.08	-0.07	-0.1
Subsidized	56	1836	0	-0.01	0	-0.04	0	0
Private	44	497	0	0.07	0.08	-0.29	-0.01	-0.01
Region:								
Center	54	1804	-0.11	-0.13	-0.11	-0.42	-0.11	-0.11
North	28	356	-0.02	-0.01	-0.07	0.04	-0.05	0.03
South	54	2181	-0.06	-0.07	-0.08	-0.04	-0.06	-0.08

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 568: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Educación_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	35	1316	-0.04	-0.01	-0.01	-0.04	0	0.03
Female	35	2365	0.01	-0.01	-0.03	0.01	-0.01	-0.03
SES:								
QA	35	932	0.04	0.02	0.01	0.05	0.05	0.03
QB	35	1202	0.03	0.05	0.03	0.03	0.03	0.04
QC	35	559	-0.01	-0.02	0.03	-0.02	0	-0.07
QD	35	295	-0.02	0.02	0.17	-0.04	-0.02	0.06
QE	31	162	0.1	0.09	0.14	0.11	0.13	0.24
Curricular Branch:								
Scientific-Humanistic	35	3194	0	0	0	0	0	-0.01
Technical-Professional	35	487	-0.05	-0.04	0	-0.09	-0.05	0.1
High School Type:								
Municipal	35	1757	-0.02	0	-0.04	-0.02	-0.02	0
Subsidized	35	1834	-0.01	-0.01	0.02	0	0	0
Private	29	90	-0.28	-0.18	-0.03	-0.24	-0.22	-0.17
Region:								
Center	32	1164	0.02	0	0.19	0	0.03	0.15
North	20	1183	-0.15	-0.16	0	-0.13	-0.21	-0.29
South	35	1334	0.04	0.01	-0.07	0.05	0.03	0.09

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 569: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Educación_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	14	839	-0.05	-0.05	-0.07	-0.05	-0.04	-0.05
Female	14	649	0.09	0.08	0.12	0.09	0.07	0.08
SES:								
QA	18	257	0	0	-0.09	0.09	0	-0.09
QB	18	398	0.01	0.02	-0.03	0	0	-0.03
QC	18	253	0.08	0.06	0.08	0.04	0.04	0.05
QD	18	149	0.09	0.09	0.2	0.06	0.11	0.08
QE	16	126	-0.21	-0.23	-0.31	-0.37	-0.22	-0.11
Curricular Branch:								
Scientific-Humanistic	18	1360	-0.03	-0.03	-0.03	-0.04	-0.03	-0.03
Technical-Professional	18	128	0.17	0.19	0.23	0.32	0.16	0.19
High School Type:								
Municipal	18	541	0.01	-0.01	-0.03	-0.02	-0.01	-0.01
Subsidized	18	742	0.02	0.03	0.04	0.05	0.03	0.04
Private	17	205	-0.1	-0.13	-0.12	0.03	-0.12	-0.09
Region:								
Center	18	704	0.08	-0.03	0	-0.03	0.03	0.11
North	11	60	-0.17	-0.09	-0.28	0.33	-0.19	0.32
South	17	724	-0.04	-0.04	0.02	-0.11	-0.06	-0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 570: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (General)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	331	0.04	0.03	0.04	0.04	0.06	0.02
Female	1	373	-0.03	-0.03	-0.03	-0.04	-0.06	-0.02
SES:								
QA	1	53	0.39	0.41	0.44	0.44	0.36	0.41
QB	1	69	-0.06	-0.1	-0.11	-0.08	-0.07	-0.15
QC	1	70	0.02	0.03	-0.14	0.01	0.04	0.07
QD	1	70	0.01	0.01	0.1	0	0.02	-0.05
QE	1	78	0.06	0.02	0	0.09	0.06	0.01
Curricular Branch:								
Scientific-Humanistic	1	695	0	0.01	0.01	0	0	0
Technical-Professional	1	9	-0.11	-0.38	-0.33	-0.15	-0.11	-0.31
High School Type:								
Municipal	1	166	-0.06	-0.08	-0.12	-0.05	-0.02	-0.13
Subsidized	1	231	0.04	0.05	-0.01	0.03	0.01	0.07
Private	1	307	-0.01	0.01	0.07	-0.01	-0.01	0.02
Region:								
Center	1	571	-0.01	-0.01	-0.04	-0.01	-0.01	-0.01
North	1	34	0.2	0.2	0.42	0.21	0.19	0.2
South	1	99	-0.03	-0.01	0.05	-0.06	-0.04	0.01

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 571: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Humanidades)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	14	586	0.01	-0.02	0.01	-0.22	0.04	0.05
Female	14	805	-0.02	0	-0.02	0.24	-0.03	-0.02
SES:								
QA	13	134	0.02	0.04	-0.02	-0.38	0	0.09
QB	14	182	-0.01	-0.02	-0.07	0.1	-0.03	-0.04
QC	14	154	-0.01	0.01	0.04	-0.27	-0.01	0.05
QD	14	135	-0.08	0	-0.03	0.1	0.01	0.05
QE	13	145	-0.09	-0.03	-0.04	-0.03	-0.03	0.03
Curricular Branch:								
Scientific-Humanistic	14	1339	0.01	0	0.01	-0.03	0.01	0.01
Technical-Professional	13	51	-0.2	-0.15	-0.15	0.32	-0.21	-0.34
High School Type:								
Municipal	13	338	0.05	0.06	0.05	0.03	0.08	0.02
Subsidized	14	600	-0.08	-0.07	-0.08	-0.09	-0.08	-0.04
Private	14	452	-0.02	-0.04	-0.04	0.08	-0.05	0.02
Region:								
Center	14	1129	-0.05	-0.04	-0.06	-0.02	-0.06	-0.07
North	12	37	-0.03	-0.1	0.04	-0.55	-0.05	0.02
South	13	224	0.01	0.08	0.03	0.46	0.03	0.23

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 572: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ingeniería_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	113	13199	0	-0.01	0.01	0	-0.01	-0.01
Female	112	3586	0.03	0.05	-0.03	0.02	0.04	0.02
SES:								
QA	113	1845	0.09	0.08	0.07	0.09	0.1	0.06
QB	112	2531	0.06	0.05	0.04	0.03	0.07	0.05
QC	112	2056	-0.05	-0.05	-0.02	-0.04	-0.05	-0.05
QD	113	1569	0.01	0.02	0.08	0.03	0.02	-0.01
QE	112	1764	0.04	0.02	0.01	0.04	0.02	0.03
Curricular Branch:								
Scientific-Humanistic	113	15289	0	-0.01	0.01	0	0	-0.02
Technical-Professional	107	1493	0.01	0.02	0.03	0.03	-0.01	0.07
High School Type:								
Municipal	113	4915	-0.01	-0.01	-0.04	-0.01	0	-0.01
Subsidized	113	6735	-0.01	-0.01	0	-0.01	-0.02	-0.03
Private	109	5132	-0.01	-0.01	0.02	-0.02	-0.03	-0.07
Region:								
Center	103	8105	0.01	0.02	-0.04	0.02	0.02	0.03
North	90	2419	-0.01	-0.03	-0.1	-0.02	-0.01	-0.06
South	104	6258	0.01	0.02	0.05	0.02	0.04	0.04

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 573: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ingeniería_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	64	3262	-0.01	-0.01	-0.02	-0.02	0	-0.02
Female	62	1608	0.01	0.03	0.1	0.02	0.02	0.02
SES:								
QA	63	750	0.01	0.02	0.06	0.01	0.01	0.06
QB	64	1203	0.02	-0.02	-0.01	0	-0.03	0.01
QC	63	846	0.01	0	0.01	0	0	0
QD	64	556	0.07	0	0.18	0.01	0.08	0
QE	60	336	0.01	-0.01	-0.11	0.06	0.02	0.02
Curricular Branch:								
Scientific-Humanistic	64	4175	0	0.01	-0.01	0	0	0
Technical-Professional	63	692	0	-0.08	-0.08	-0.04	-0.05	-0.07
High School Type:								
Municipal	64	1861	-0.02	-0.02	-0.06	0	-0.01	-0.03
Subsidized	64	2545	0.03	0.01	0.06	0	0.01	-0.01
Private	61	461	-0.09	-0.04	-0.03	-0.1	-0.03	0.02
Region:								
Center	61	2878	0.02	0.01	-0.06	0.01	0	0.01
North	39	297	0.1	0.12	-0.11	0.1	0.1	0.12
South	62	1692	-0.06	-0.03	-0.17	-0.06	-0.02	-0.12

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 574: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Ingeniería_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	9	329	0.07	0.08	-0.05	0.08	0.13	0.12
Female	9	310	-0.07	-0.07	0.03	-0.07	-0.11	-0.1
SES:								
QA	9	78	0.04	0.02	-0.03	-0.04	0	-0.01
QB	8	77	-0.11	-0.08	-0.06	-0.1	-0.08	-0.14
QC	9	85	0.11	0.1	-0.36	0.1	0.05	0.15
QD	9	62	-0.27	-0.32	-0.13	-0.27	-0.26	-0.32
QE	9	73	-0.03	-0.02	0.52	0.01	0.03	0.06
Curricular Branch:								
Scientific-Humanistic	9	622	0	0.01	0.02	0	0.01	0.01
Technical-Professional	5	17	-0.17	-0.16	-0.54	-0.23	-0.24	-0.38
High School Type:								
Municipal	8	187	-0.07	-0.06	-0.25	-0.08	-0.07	-0.09
Subsidized	9	273	0.05	0.04	-0.13	0.07	0.06	0.05
Private	9	179	-0.12	-0.08	0.39	-0.08	-0.07	-0.03
Region:								
Center	9	273	-0.24	-0.2	-1.17	-0.25	-0.19	-0.17
North	8	84	-0.34	-0.39	-0.36	-0.28	-0.41	-0.41
South	9	282	-0.17	-0.16	0.05	-0.17	-0.16	-0.23

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 575: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Mar)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	19	473	0.02	0.04	0.13	0.01	0.03	0.1
Female	19	536	-0.05	-0.08	0	-0.05	-0.06	-0.14
SES:								
QA	19	176	0.21	0.24	-0.04	0.27	0.12	0.1
QB	19	231	0	-0.05	0.01	-0.06	-0.01	-0.02
QC	19	168	0.08	0.11	-0.12	0.06	0.05	0.04
QD	18	108	-0.17	-0.1	0.65	-0.11	-0.1	-0.14
QE	16	91	0.03	-0.14	-0.18	0.05	0.06	0.14
Curricular Branch:								
Scientific-Humanistic	19	932	0	0.01	0	0.01	0	0.01
Technical-Professional	18	77	0.08	0.06	0.06	-0.12	0.07	0.13
High School Type:								
Municipal	19	438	0.06	0.07	-0.06	0.06	0.07	-0.08
Subsidized	19	435	-0.05	-0.06	-0.04	-0.06	-0.05	0.03
Private	18	136	-0.19	-0.24	-0.03	-0.14	-0.26	-0.13
Region:								
Center	17	297	0.06	0.04	0.06	0.04	0.08	0
North	13	195	-0.06	0.09	0.38	-0.08	0.09	0.14
South	17	517	-0.13	-0.16	-0.04	-0.12	-0.13	-0.13

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 576: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Periodismo)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	16	748	-0.04	-0.05	-0.05	-0.15	0	-0.04
Female	16	1247	0.04	0.07	0.06	0.2	0.03	0.03
SES:								
QA	16	225	0	0.05	0.01	0.08	0.03	-0.01
QB	15	192	0.13	0.16	0.13	-0.18	0.16	0.1
QC	15	235	-0.11	-0.09	-0.1	-0.34	-0.11	-0.15
QD	15	206	-0.01	0	0.01	0.21	0.02	-0.08
QE	16	280	-0.04	-0.06	-0.02	0.28	-0.06	-0.05
Curricular Branch:								
Scientific-Humanistic	16	1906	0.01	0.01	0.01	0.03	0.01	-0.01
Technical-Professional	14	89	-0.42	-0.47	-0.39	-0.63	-0.43	-0.44
High School Type:								
Municipal	14	374	-0.15	-0.15	-0.16	-0.19	-0.12	-0.19
Subsidized	16	777	0.4	0.4	0.41	0.11	0.37	0.35
Private	16	844	-0.04	-0.03	-0.02	0.13	-0.03	-0.02
Region:								
Center	16	1227	-0.06	-0.08	-0.08	-0.58	-0.07	-0.04
North	14	205	0.03	0.3	0.1	-0.12	0.22	0.07
South	16	563	0.14	0.17	0.17	-0.04	0.14	0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 577: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Salud_1)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	27	2230	-0.01	-0.02	-0.03	-0.03	0	-0.01
Female	27	2209	0	0.01	-0.01	0.02	-0.01	0.01
SES:								
QA	27	346	0.12	0.09	-0.07	0.13	0.1	0.07
QB	26	356	0.11	0.11	0	0.09	0.07	0.03
QC	27	368	-0.09	-0.1	0.14	-0.09	-0.09	-0.06
QD	27	395	-0.04	-0.05	-0.17	-0.02	-0.05	-0.04
QE	27	547	0.01	0.02	0.31	0.01	0.01	0.04
Curricular Branch:								
Scientific-Humanistic	27	4409	0	0	-0.02	0	-0.01	0
Technical-Professional	18	30	-0.17	-0.1	0.24	-0.13	-0.12	0.02
High School Type:								
Municipal	27	807	0.05	0.03	0.02	0.05	0.04	0.07
Subsidized	27	1537	-0.01	-0.01	0.04	0	-0.02	-0.01
Private	27	2095	0	0.01	-0.03	0	0.01	0.02
Region:								
Center	27	2049	0.13	0.12	0.1	0.1	0.15	0.06
North	23	637	0	-0.01	0.08	0	-0.02	-0.09
South	27	1753	-0.03	-0.03	0.11	-0.02	-0.03	-0.07

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 578: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Salud_2)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	56	1659	-0.06	-0.06	0.11	-0.07	-0.04	-0.03
Female	56	5375	0.02	0.01	-0.02	0.02	0	0.01
SES:								
QA	56	982	0.05	0.02	0.01	0.03	0.02	0.01
QB	56	1485	-0.01	0.01	-0.02	0.01	-0.02	-0.03
QC	56	1205	0.05	0.05	-0.03	0.05	0.05	0.05
QD	56	832	0.02	-0.01	-0.17	0.02	0	0.02
QE	56	688	-0.01	0	0.11	-0.02	0.02	0.01
Curricular Branch:								
Scientific-Humanistic	56	6738	0	0	-0.01	0	0	0
Technical-Professional	53	296	0.04	0.02	0.16	0.05	0.02	0
High School Type:								
Municipal	56	2320	0	0	0	0	0	0
Subsidized	56	3526	0	0	0.02	0	0	-0.02
Private	53	1188	-0.07	-0.07	-0.22	-0.08	-0.04	-0.03
Region:								
Center	52	3215	0.15	0.16	-0.1	0.17	0.16	0.01
North	39	733	-0.15	-0.11	0.08	-0.06	-0.13	-0.11
South	54	3086	-0.05	-0.05	-0.18	-0.03	-0.05	-0.05

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 579: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Salud_3)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	24	868	-0.1	-0.1	-0.07	-0.11	-0.07	-0.14
Female	24	2192	0.01	0.01	-0.02	0.01	-0.01	0.02
SES:								
QA	24	408	-0.03	-0.02	-0.02	0	0.01	0
QB	24	623	-0.04	-0.05	0.08	-0.06	-0.06	-0.07
QC	24	502	0.07	0.07	0.03	0.07	0.09	0.08
QD	24	380	-0.01	-0.01	-0.14	0	0.01	0.03
QE	24	277	0.07	0.08	-0.36	0.07	0.08	-0.06
Curricular Branch:								
Scientific-Humanistic	24	2943	0	0.01	0	0	0	0.01
Technical-Professional	22	115	-0.1	-0.17	0.09	-0.12	-0.08	-0.09
High School Type:								
Municipal	24	1045	0.03	0.02	0.08	0.02	0.03	-0.01
Subsidized	24	1552	0.01	0.01	-0.07	0.01	0	-0.02
Private	24	461	-0.07	-0.06	-0.17	-0.07	-0.01	0.02
Region:								
Center	22	1479	0.11	0.12	-0.13	0.11	0.11	0.07
North	20	208	0.07	0.1	0.39	0.05	0.1	-0.1
South	23	1371	-0.01	-0.01	0.01	-0.02	-0.04	-0.03

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 580: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Técnico_Administración)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	7	283	0.03	0	-0.02	0.24	0.03	0.01
Female	7	438	-0.02	0.02	0.01	-0.2	-0.02	0
SES:								
QA	7	129	-0.04	-0.05	-0.05	-0.18	-0.05	-0.14
QB	7	183	0.04	0.03	0.06	0.12	0.06	0.02
QC	7	138	-0.05	-0.54	-0.07	0.01	-0.1	0.14
QD	7	89	-0.03	0	0.07	-0.1	0.02	-0.02
QE	7	57	0.13	0.18	0.03	0.24	0.18	0.17
Curricular Branch:								
Scientific-Humanistic	7	532	-0.03	-0.04	-0.03	0.07	-0.02	0
Technical-Professional	7	189	0.07	0.07	0.09	-0.19	0.02	-0.02
High School Type:								
Municipal	7	246	-0.05	-0.05	-0.03	-0.12	-0.04	-0.12
Subsidized	7	415	0.04	-0.01	0	0.11	0.02	0.04
Private	7	60	0.09	0.16	0.19	0.01	0.15	0.18
Region:								
Center	7	544	0.04	0.02	-0.02	-0.24	0.01	0
North	6	94	-0.32	-0.43	-0.32	-0.15	-0.43	-0.67
South	7	83	0.01	0.07	0.05	0.09	0.07	0.08

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 581: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Técnico_Agro)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	12	0	0	0.16	0.02	0.07	0.1
Female	1	8	0.01	-0.01	-0.31	-0.03	-0.13	-0.16
SES:								
QA	1	6	0.08	0.11	-0.16	0.1	0	-0.02
QB	1	5	-0.06	-0.1	-0.24	-0.12	-0.07	-0.13
QC	1	5	-0.1	-0.06	0.37	-0.02	0.03	0.2
QD	1	1	0.27	0.28	0.38	0.33	0.69	-0.45
QE	-	-	-	-	-	-	-	-
Curricular Branch:								
Scientific-Humanistic	1	13	-0.1	-0.07	-0.08	-0.08	-0.06	0.27
Technical-Professional	1	7	0.21	0.17	0.28	0.19	0.19	-0.51
High School Type:								
Municipal	1	12	-0.02	0	0.09	-0.02	0	-0.12
Subsidized	1	8	0.08	0.05	-0.2	0.1	0.1	0.17
Private	-	-	-	-	-	-	-	-
Region:								
Center	1	1	0.38	0.37	-	0.33	0.32	1.84
North	-	-	-	-	-	-	-	-
South	1	19	-0.02	-0.02	-0.01	-0.01	-0.01	-0.1

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 582: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Técnico_Ciencias)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	4	87	-0.03	0.07	-0.26	-0.06	-0.06	-0.09
Female	4	112	0.03	0.01	0.23	0.04	0.03	0.04
SES:								
QA	4	37	0.14	0.08	0.22	0.12	0.17	0.24
QB	4	66	0.23	0.28	0.31	0.14	0.2	0.12
QC	4	32	-0.02	-0.01	0.21	-0.02	-0.02	0.12
QD	4	20	-0.07	0.35	0.28	-0.21	-0.12	-0.42
QE	3	12	-0.54	-0.35	0.23	-0.4	-0.47	-0.26
Curricular Branch:								
Scientific-Humanistic	4	185	-0.01	0.05	0	0	0	0
Technical-Professional	4	14	-0.32	-0.83	-	-0.35	-0.36	-0.78
High School Type:								
Municipal	4	122	0.05	-0.01	-0.08	0.13	0.03	0.01
Subsidized	4	67	-0.13	0.15	0.29	-0.23	-0.11	-0.07
Private	3	10	0.25	-0.33	-	0.29	0.25	0.04
Region:								
Center	2	44	-0.37	-0.16	-	-0.53	-0.14	-0.35
North	3	3	0.57	0.65	-	0.65	0.58	0.81
South	3	152	-0.09	0.02	0	-0.07	-0.23	0.11

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 583: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Técnico_Diseño)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	3	79	-0.02	0	0.01	0.04	-0.02	0.02
Female	3	63	0.07	0.04	0.09	0.03	0.12	0.08
SES:								
QA	3	23	0.44	0.26	0.38	0.35	0.32	0.48
QB	3	41	0.04	0.06	0.12	0.05	-0.04	0.01
QC	3	26	-0.11	0.07	-0.16	-0.48	-0.09	0.09
QD	3	21	-0.14	-0.22	-0.37	0.37	-0.11	-0.12
QE	3	7	0	0.24	0.19	0.5	0.12	0.06
Curricular Branch:								
Scientific-Humanistic	3	121	-0.09	-0.03	-0.05	-0.01	-0.06	-0.02
Technical-Professional	3	21	0.41	0.13	0.26	0.03	0.3	0.05
High School Type:								
Municipal	3	55	-0.01	-0.15	-0.12	-0.05	-0.09	-0.08
Subsidized	3	81	0.03	0.08	0.15	0.07	0.11	0.03
Private	3	6	-0.29	-0.05	-0.28	0.21	-0.31	0.02
Region:								
Center	3	137	0	0	-0.01	-0.01	0	0
North	-	-	0.26	-	0.46	0.49	0.36	-
South	2	5	-0.13	-0.18	-0.04	-	0.12	-0.1

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 584: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Técnico_Educación)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	1	26	0.03	0.09	0.06	0.05	0.09	0.17
Female	1	23	-0.07	-0.15	-0.15	-0.11	-0.18	-0.2
SES:								
QA	1	8	0.01	-0.09	0.21	-0.13	-0.05	0.1
QB	1	20	0.07	0.17	-0.13	0.11	0.06	0.12
QC	1	8	0.13	0	0.03	0.22	0.13	-0.06
QD	1	6	-0.44	-0.44	-0.17	-0.66	-0.4	-0.51
QE	1	2	0.18	0.51	0.27	0.61	0.61	-0.4
Curricular Branch:								
Scientific-Humanistic	1	41	-0.1	-0.07	-0.12	-0.08	-0.1	-0.09
Technical-Professional	1	8	0.41	0.3	0.47	0.37	0.36	0.46
High School Type:								
Municipal	1	13	0.24	0.17	0.19	0.26	0.27	0.17
Subsidized	1	35	-0.1	-0.04	-0.15	-0.03	-0.12	-0.05
Private	1	1	-0.34	-0.46	-0.01	-0.61	-0.36	-0.58
Region:								
Center	1	39	-0.02	-0.02	-0.07	-0.04	-0.01	-0.06
North	1	5	-0.05	0.12	0.49	-0.04	0.1	0.04
South	1	5	0.19	0.19	0.18	0.32	0.02	0.41

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 585: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Técnico_Idioma)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	7	288	-0.07	-0.01	-0.09	0.03	-0.01	-0.06
Female	7	513	0.02	-0.01	0.03	-0.02	-0.01	0.01
SES:								
QA	7	135	0.19	0.27	0.18	-0.14	0.22	0.02
QB	7	225	-0.09	-0.09	-0.11	0.1	-0.09	0.05
QC	7	123	-0.07	-0.06	-0.06	0.68	-0.03	-0.06
QD	7	87	-0.07	-0.19	-0.15	-0.4	-0.07	0.07
QE	5	55	0.05	0.15	0.12	-0.13	0.18	0.14
Curricular Branch:								
Scientific-Humanistic	7	664	-0.03	-0.02	-0.03	0.01	0	0.04
Technical-Professional	7	137	0.05	0.06	0.05	-0.06	-0.01	-0.02
High School Type:								
Municipal	7	297	-0.04	-0.02	-0.05	-0.05	-0.04	-0.09
Subsidized	7	451	-0.01	-0.04	0.01	0.05	-0.02	0.06
Private	7	53	0.28	0.41	0.23	0.22	0.47	0.09
Region:								
Center	7	208	0.18	0.17	0.17	0.05	0.22	0.02
North	7	361	0.29	0.23	0.29	0.1	0.19	0.27
South	7	232	-0.11	-0.1	-0.14	0.06	-0.08	-0.02

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 586: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Técnico_Ingeniería)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	60	2755	-0.01	0	-0.01	-0.02	0	0.01
Female	57	725	-0.08	-0.1	-0.11	-0.05	-0.1	-0.07
SES:								
QA	60	616	-0.01	-0.02	-0.05	0.03	-0.03	-0.01
QB	59	985	0.06	0.06	0.02	0.06	0.06	0.04
QC	60	634	-0.01	0.01	0.09	0.01	0	0.08
QD	57	341	-0.02	0.01	0.05	-0.06	-0.04	-0.06
QE	52	232	0.13	0.16	0.06	0.14	0.17	0.21
Curricular Branch:								
Scientific-Humanistic	60	2599	-0.02	0	-0.01	-0.02	-0.01	0
Technical-Professional	59	881	0.08	0.05	0.01	0.1	0.05	0.08
High School Type:								
Municipal	60	1460	-0.02	-0.02	-0.05	-0.01	-0.03	-0.04
Subsidized	59	1806	0.05	0.05	0.12	0.03	0.06	0.08
Private	54	214	-0.27	-0.25	-0.2	-0.19	-0.15	-0.03
Region:								
Center	55	2268	0.05	0.03	0.06	0.02	0.01	0.02
North	41	308	-0.01	-0.01	-0.3	-0.04	-0.1	-0.04
South	53	904	0.04	0.07	-0.02	0.06	0.05	0.06

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Table 587: Average Over Prediction (-) and Under Prediction (+) of University Completion by Predictor Measure and Subgroups (Veterinaria)

	Samples		PSU Tests				High School	
Sub-groups	Career	Student	Math	Language	History	Science	NEM	Ranking
Gender:								
Male	5	478	-0.02	-0.02	0.04	-0.03	0.02	-0.01
Female	5	645	0.02	0.02	-0.02	0.02	-0.01	0
SES:								
QA	5	132	0.01	0.03	0.17	-0.02	0.02	-0.02
QB	5	177	-0.06	-0.06	-0.25	-0.03	-0.05	-0.04
QC	5	152	-0.02	0	-0.02	-0.01	-0.01	0
QD	5	142	0.19	0.18	0.39	0.16	0.19	0.15
QE	5	125	0.09	0.08	0.33	0.07	0.09	0.07
Curricular Branch:								
Scientific-Humanistic	5	1079	0	0	0.01	0	0	-0.01
Technical-Professional	5	43	0.13	0.16	0.05	0.12	0.1	0.13
High School Type:								
Municipal	5	323	-0.04	-0.03	-0.15	-0.03	-0.02	-0.02
Subsidized	5	508	0	0	0.07	0	-0.02	-0.04
Private	5	291	0.07	0.07	0.29	0.07	0.11	0.12
Region:								
Center	5	522	-0.03	-0.05	0.13	-0.05	-0.03	-0.05
North	5	70	-0.15	-0.2	-0.31	-0.12	-0.19	-0.2
South	5	530	0.06	0.08	0	0.06	0.07	0.02

(Note: Mean residuals based on standardized within-career first year university grade point average are provided. Negative values indicate over prediction. Positive values indicate under prediction. Values are computed by subtracting predicted from observed university completion standing. Prediction equations are estimated within careers.)

Appendix Y. *Revisión de Marcos Teóricos de Evaluación para PSU*

In response to a July 2009 request from DEMRE, the Curriculum Unit analyzed the PSU assessment frameworks in relation to the curriculum. See the document, *Revisión de Marcos Teóricos de Evaluación para PSU*, included in the following pages of this appendix. This review concluded that the PSU does not achieve an appropriate curricular reference because it is concentrated upon the Minimum Obligatory Contents (CMO) and not upon the Fundamental Objectives (OF), which are the nucleus of the National Curriculum. One of the recommendations of this report was to strengthen the relation between the Curriculum Unit and DEMRE. The joint work would allow for a more faithful interpretation of the National Curriculum and an improved prioritization of the educational objectives of the National Curriculum that are targeted by the PSU.

The appended document is a review of the PSU by the Curriculum Unit of MINEDUC.

Please note: The pagination visible on the following pages is that of the original report from the Curriculum Unit.

REVISIÓN DE MARCOS TEÓRICOS DE EVALUACIÓN PARA PSU

Unidad de Currículum y Evaluación
Ministerio de educación

Septiembre 2009

Índice

Introducción

Con la finalidad de recibir observaciones y retroalimentación sobre el proceso de elaboración de la PSU, el DEMRE solicitó a la Unidad de Currículum y Evaluación del Mineduc (UCE) examinar los marcos a partir de los cuales esta prueba es diseñada. Dicha solicitud requería la revisión de los marcos de las pruebas de Matemática (PSU-M), Lenguaje y Comunicación (PSU-L), Ciencias (PSU-Ciencias), e Historia y Ciencias Sociales (PSU-HyCS).

De acuerdo a la petición del DEMRE, se solicitó que el examen de los marcos de evaluación tomara en consideración una serie de puntos: Claridad general de la propuesta, metodología de análisis del currículum, noción de referencia curricular, relación con el Marco Curricular, noción de habilidad cognitiva, criterios técnicos para la elaboración de ítems, y supuestos teóricos. Para efectos de este análisis, estos distintos puntos se discuten en función de un eje principal: el grado en que estos marcos evaluativos favorecen el alineamiento de la prueba con el currículum nacional para la Enseñanza Media, esto es, la medida en que garantizan la elaboración de una prueba que efectivamente evalúe los aprendizajes que el currículum prescribe.

La presentación del análisis llevado a cabo se organiza en el presente documento en dos partes. La primera parte ofrece una visión general referida al conjunto de los marcos evaluativos. Por medio de esta visión general se plantean aquellas observaciones de mayor relevancia y que a la vez son pertinentes para los cuatro marcos analizados. Junto con estas observaciones, en esta primera parte se sugieren también algunas líneas de trabajo tendientes a potenciar el alineamiento curricular de los marcos evaluativos. En la segunda parte del informe se presenta el análisis específico realizado para cada uno de los marcos evaluativos. Esta parte se organiza en cuatro capítulos, cada uno de ellos destinado a presentar las observaciones sobre el marco evaluativo de una prueba en particular.

Antes de abordar directamente el análisis de acuerdo a los lineamientos arriba señalados, resulta necesario indicar que las observaciones refieren a los marcos evaluativos elaborados para cada prueba, y no a las pruebas en sí mismas. Un análisis de estas últimas requeriría, por una parte, un examen detallado de las preguntas que se realizan en la PSU, aspecto que no constituye el foco de atención en el presente documento. Por otra parte, un análisis de las pruebas requeriría también un examen de los aspectos métricos involucrados en su construcción, a los cuales esta unidad no ha tenido acceso.

Finalmente, cabe señalar que los análisis específicos de cada uno de los marcos evaluativos fueron llevados a cabo por diferentes equipos de especialistas en cada una de las áreas evaluadas. Si bien en cada caso esta tarea se articuló de acuerdo a una misma lógica, cada uno de estos equipos utilizó una estructura distinta para organizar sus respectivos informes, en consideración de los argumentos y observaciones propios de cada área. Junto con ello, algunos de los equipos consideraron también algunos elementos complementarios a los marcos propiamente tales (como algunas preguntas de la prueba o la tabla de conversión de puntaje), con la finalidad de enriquecer las observaciones relativas a la elaboración de la prueba.

Parte I. Reporte Integrado

Capítulo 1. Marcos Evaluativos y Alineamiento al Currículum Nacional

El presente capítulo presenta las principales observaciones realizadas a partir de una mirada de conjunto sobre los marcos evaluativos para la PSU. En éste se integran aquellas observaciones que de alguna u otra forma se expresan en el análisis de cada uno de estos marcos, y que en su conjunto son consideradas de especial significancia desde el punto de vista curricular. Como se señaló en la introducción de este informe, en este análisis se abordan diversos aspectos de los marcos evaluativos en función de la relación que éstos permiten establecer entre la prueba y el currículum nacional[1].

El capítulo se organiza en tres secciones. En la primera de ellas se presentan los principales antecedentes que se tienen cuenta para realizar el presente análisis. En la segunda sección se dan a conocer las principales observaciones realizadas en torno a los marcos evaluativos y el alineamiento al currículum que éstos favorecen. Finalmente, en la tercera sección se proponen algunas líneas de acción orientadas a aportar al desarrollo y fortalecimiento de los marcos evaluativos.

1.1 Antecedentes

La elaboración de los marcos evaluativos para la PSU supone una contribución para el desarrollo de la prueba en varios sentidos. En primer término, debe tenerse en cuenta la decisión de formular un documento de este tipo. Al realizar estos marcos, se sigue la línea adoptada internacionalmente para la elaboración de pruebas de aprendizaje. La formulación de éstos tiene el valor de sistematizar y dar a conocer los conceptos básicos, los procedimientos y las categorías que subyacen a los instrumentos de evaluación. De tal forma, favorecen el desarrollo de un lenguaje común, así como la comprensión de la prueba y los resultados obtenidos a través de ésta.

Junto con ello, cabe destacar el hecho que se considere como parte constitutiva de los marcos evaluativos para la PSU el análisis del currículum de la Enseñanza Media. Dado que esta prueba se define como un instrumento referido a las bases curriculares, dicho análisis constituye un antecedente fundamental. Al respecto cabe destacar que estos marcos no sólo realizan un análisis del currículum en los cuatro sectores evaluados, sino que también establecen una metodología sistemática para llevar a cabo este análisis curricular. De esta forma, la transferencia del currículum a la PSU se sostiene sobre un procedimiento que busca asegurar condiciones de precisión y rigurosidad.

[1] Dado que la recientemente promulgada Ley General de Educación (LGE) se refiere al currículum nacional bajo el concepto de "bases curriculares", el presente informe utiliza esta terminología en lugar del concepto de "marco curricular" comúnmente utilizado para referirse al documento de base del currículum oficial.

Las observaciones que se presentan en este informe parten del reconocimiento que los marcos evaluativos son por excelencia el recurso para lograr el alineamiento de la prueba con el curriculum. En consecuencia, el presente trabajo aborda los marcos de evaluación existentes con el propósito de realizar recomendaciones que permitan una mejor alineación de la PSU con el curriculum nacional.

Antes de presentar estas observaciones resulta importante considerar algunos de los antecedentes sobre los cuales se sustentan estas recomendaciones. En primer lugar se expondrán las razones por las cuales se decide organizar el presente análisis en función del alineamiento curricular. En segundo lugar, se presentan las categorías en base a las cuales se analizó la sujeción de los marcos evaluativos al currículum.

a) Alineamiento curricular como foco de análisis

La opción de articular el presente análisis en función del alineamiento al currículum responde a dos motivos. Por una parte, este es uno de los principales ámbitos de competencia sobre los que corresponde pronunciarse a la UCE, toda vez que es el organismo especializado en la tarea ministerial de realizar desarrollo curricular. En consecuencia es desde esta instancia que cabe plantear el juicio respecto del grado en que el currículum se logra ver reflejado en los marcos evaluativos de la PSU. Por otra parte, se adopta el alineamiento como categoría central debido a que este es una finalidad declarada de la PSU en tanto instrumento de selección universitaria. Se debe recordar que uno de los propósitos centrales de esta nueva prueba era transformar en relevantes para los estudiantes y establecimientos educacionales la trayectoria de todos los años de la enseñanza media, cuestión que se observaba, no ocurría en el caso de la PAA.

Como se señala en los mismos marcos evaluativos al referirse a este punto, en el marco de los acuerdos logrados en la Mesa Escolar en el año 2002, la decisión de cambiar la prueba de selección universitaria es concebida en función de un doble propósito. Por una parte, responde a la necesidad de contar con un predictor del desempeño académico que sirva como herramienta para el proceso de selección universitaria. Por otra parte, este cambio responde al propósito de vincular la prueba de selección universitaria con el logro de los aprendizajes establecidos para la Educación Media, reforzando de esta forma la experiencia formativa de la educación escolar. En consecuencia, esta evaluación se transformó en un instrumento cuya función no se agota en la tarea de selección, sino que a la vez se erige como un medio para dotar de una mayor relevancia y promover los aprendizajes a desarrollar en los últimos cuatro años de la formación escolar.

Es en función de este último propósito que el logro de un alineamiento efectivo de la prueba con el currículum resulta un elemento central. Un instrumento que no garantice la evaluación efectiva de aquellos aprendizajes establecidos en el currículum supone el riesgo de vincular la selección universitaria a un currículum paralelo, esto es, a la evaluación de un conjunto de aprendizajes distintos de aquellos que se definen para el sistema escolar, introduciendo de esta forma un elemento que altera o distorsiona la promoción de los aprendizajes establecidos para la Educación Media a través de las bases curriculares.

b) Condiciones para el alineamiento curricular

En tanto la sujeción al currículum constituye una de las propiedades esperadas de la PSU, resulta necesario considerar como primer antecedente algunas características básicas del currículum, a partir de las cuales se desprenden implicancias importantes para efectos de lograr el alineamiento de la prueba al mismo.

El currículum está conformado por diversos elementos y cada uno de ellos es importante para comprender los aprendizajes que se promueven. Estos elementos son:

- La introducción general del currículum nacional, por medio de la cual se da a conocer las orientaciones y opciones básicas que subyacen a la totalidad del currículum.
- Los Objetivos Fundamentales Transversales, que expresan propósitos de aprendizaje y formativos que permean los distintos sectores, y que en consecuencia, forman también parte constitutiva de los logros a alcanzar en éstos.
- Los sectores de aprendizaje, que especifican aquello a ser aprendido en las distintas áreas de conocimiento. Por una parte, la introducción de cada uno de estos sectores expresa la naturaleza y el carácter general de los aprendizajes a desarrollar. Por otra parte, los Objetivos Fundamentales (OF) y Contenidos Mínimos Obligatorios (CMO) de cada sector concretizan estos aprendizajes, así como de las orientaciones generales adoptadas en el currículum, en términos de aquello a ir logrando año a año en cada uno de los niveles educativos.

Dado que los OF y CMO concretizan la propuesta curricular, ambos constituyen referentes importantes para la elaboración de la prueba. Sin embargo, cabe señalar que estos dos elementos no son equivalentes en términos de su significancia en el currículum. Este último está elaborado de acuerdo a un diseño que contempla como principal categoría los OF. Éstos señalan los aprendizajes a desarrollar. De acuerdo a lo establecido en el currículum nacional, los OF definen "las competencias o capacidades que los alumnos y alumnas deben lograr al finalizar los distintos niveles de la Educación Media y que constituyen el fin que orienta al conjunto del proceso de enseñanza-aprendizaje"[2].

En cuanto a los CMO, si bien son una categoría importante, no constituyen el elemento articulador del currículum. En rigor, éstos son incorporados en función de los OF. De acuerdo a la definición que se presenta en el currículum, los CMO "son los conocimientos específicos y prácticas para lograr habilidades y actitudes que los establecimientos deben obligatoriamente enseñar, cultivar y promover *para cumplir los objetivos fundamentales establecidos para cada nivel*"[3]. Esto implica que la presencia de los CMO en el currículum no se justifica en sí misma, sino que responde a su rol como medio para el desarrollo de los OF.

[2] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 7).
[3] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 8). Énfasis añadido.

La distinción arriba señalada tiene una implicancia fundamental para efectos de concebir el alineamiento curricular de los marcos evaluativos. Implica que *el alineamiento de la PSU al currículum va a estar dado principalmente por el grado en que aquello que se evalúa se corresponde con los aprendizajes que se definen en los OF*. Mientras, la evaluación sólo de contenidos no garantiza dicho alineamiento.

Esta es una distinción que se aplica al currículum vigente, pero que resulta especialmente significativa en función de los cambios y nuevas regulaciones que afectan el currículum. En el ajuste curricular recientemente aprobado por el Consejo Nacional de Educación se reafirma el lugar central de los OF como definición de los aprendizajes a desarrollar, así como el carácter instrumental de los CMO en relación a estos aprendizajes.

"Si los Objetivos Fundamentales están formulados desde la perspectiva del aprendizaje que cada alumno y alumna debe lograr, los CMO lo están desde la perspectiva de lo que cada docente debe obligatoriamente enseñar, cultivar y promover en el aula y en el espacio mayor del establecimiento, para desarrollar dichos aprendizajes"[4].

Por otra parte, la creciente importancia que adoptan los objetivos de aprendizajes en relación a los contenidos queda de manifiesto también en la recientemente promulgada Ley General de Educación (LGE). Este cuerpo legal no hace mención de los contenidos como parte constitutiva de las bases curriculares para la educación escolar, sino sólo a los objetivos de aprendizaje en función de los cuales estas bases se articulan[5]. De tal forma, se decanta en esta ley una concepción de currículum de acuerdo a la cual la categoría central del mismo está constituida por tales objetivos.

Estos antecedentes apuntan al hecho que si bien actualmente los OF merecen ser considerados como el referente central para lograr la sujeción al currículum, esta es una situación que se verá reforzada en el futuro.

1.2 Observaciones sobre los marcos evaluativos

El análisis de los marcos evaluativos arroja como resultado que éstos no se encuentran suficientemente articulados a los aprendizajes que los OF intencionan, dándose un amplio espacio para su mejora. Esto implica que se puede potenciar la referencia curricular de la PSU, lo que resulta especialmente importante en atención a la reciente entrada en vigencia de la LGE. El análisis específico sobre este punto para cada uno de los marcos evaluativos se presenta en la segunda parte de este documento. En esta sección se dan a conocer las principales observaciones que son extensivas a todos ellos.

Las observaciones relativas al alineamiento de los marcos evaluativos con el currículum se organizan en función de tres aspectos de dichos marcos: a) las categorías a través de las cuales se operacionaliza la noción de referencia curricular; b) la manera en que se analizan e incorporan los OF en los marcos evaluativos, y; c) los criterios utilizados para priorizar y seleccionar aquello a ser evaluado en la PSU.

[4] Mineduc (2009) "Propuesta de ajuste curricular" (pág. 10). Disponible en http://www.curriculum-mineduc.cl/ayuda/docs/ajuste-curricular-2/Capitulos_Introductorios.pdf.
[5] Ver Artículo 31 de la LGE

a) Operacionalización de la noción de referencia curricular

Si bien en los marcos evaluativos se hace mención a los OF al momento de establecer la noción operativa de referencia curricular, estos objetivos no son utilizados como el elemento central en función del cual se busca establecer el alineamiento con el currículum. La referencia curricular es operacionalizada a través de la consideración tanto de los CMO como de las habilidades que se desprenden de los OF, sin que se asuma de manera directa las distinciones jerárquicas entre ambas, y por lo tanto, sin un claro reconocimiento de la subordinación o carácter instrumental de los CMO en relación a los OF. La ausencia de un reconocimiento directo de esta condición se aprecia en la forma en que se define la noción operativa de referencia curricular:

"Esta relación [entre el instrumento de evaluación y el Marco Curricular vigente] se hace visible al considerar que cada uno de los ítems de la prueba está **basado** en uno de los CMO, así como en una **habilidad cognitiva**. Las habilidades cognitivas a su vez, se han recogido en parte del constructo teórico anterior de la PAA-V, así como de las *acciones pedagógicas* que informan los **Objetivos Fundamentales** del subsector"[6]

Cabe señalar que la insuficiente visión de los OF como la categoría central para el alineamiento de la prueba con el currículum, no deriva exclusivamente de las decisiones técnicas adoptadas por el DEMRE, sino también de instancias de acuerdo y decisión previas, en algunas de las cuales hubo participación directa del Mineduc. Específicamente, esta es una condición que deriva en parte de los acuerdos adoptados en el Consejo de Rectores, y también en la Mesa Escolar durante el año 2002. Los acuerdos a los que se llega por medio de estas instancias aluden a la referencia curricular del nuevo sistema de selección universitaria sin señalar de manera explícita a los OF como la categoría central del currículum nacional. A la vez, en estos acuerdos se alude a este currículum principalmente a través de la categoría genérica de "contenidos", en función de la cual se tiende a conferir especial atención a los CMO. Este punto se aprecia en el señalamiento que en los mismos marcos evaluativos se realiza respecto de las decisiones adoptadas en las instancias arriba señaladas:

"La decisión de referir los instrumentos de evaluación con vistas a seleccionar postulantes a las universidades del Consejo de Rectores en los Contenidos Mínimos Obligatorios (CMO) del Marco Curricular del plan de formación general vigente, tuvo por finalidad establecer los criterios basales en los que fijar la elaboración de dichos instrumentos"[7]

Esta focalización en los CMO (y la relativa invisibilidad de los OF) se aprecia también en la forma de acuerdo a la cual se establece la implementación gradual de la evaluación del currículum. Debido al carácter reciente de la Reforma Curricular al momento de decidir la elaboración de la PSU, se definió un proceso de inclusión progresiva del currículum. Este comenzó con la inclusión parcial del mismo el año 2003, para ir aumentando en los años siguientes, de manera tal de considerar la totalidad del currículum para la aplicación realizada el año 2005. El punto significativo para efectos de este documento consiste en que ésta es una progresión en términos de la

[6] DEMRE (2008) Marco Teórico Prueba de Lenguaje y Comunicación (pág. 85); también en DEMRE (2006) Marco Teórico de la Prueba de Matemática (pág. 56).
[7] DEMRE (2008) Marco Teórico Prueba de Lenguaje y Comunicación (pág. 84); también en DEMRE (2006) Marco Teórico de la Prueba de Matemática (pág. 55).

una creciente inclusión de "contenidos", como se señala en los marcos evaluativos de la prueba[8], no así de los OF.

Lo anterior permite visualizar algunos de los antecedentes que explican porqué los OF no son considerados como la categoría central al momento de establecer la relación entre la PSU y el currículum. No obstante, este es un punto necesario de perfeccionar para en el futuro asegurar la referencia curricular de este instrumento. En tanto la referencia a los OF no logre ser rescatada, se corre el riesgo que los CMO sean considerados en sentidos diferentes a los prescritos por el curriculum, disminuyendo el alineamiento curricular de la prueba. Por ejemplo, se pueden preguntar detalles irrelevantes sobre un contenido determinado y no evaluar el entendimiento que se busca favorecer, y que se especifica a través de los OF.

A modo de ilustración, en el documento "Análisis Marco Curricular PSU Ciencias-Química" (pág. 70), se observa que se formulan preguntas como la siguiente:

> ¿Cuál es la geometría de la molécula de CS_2?
>
> A) Angular.
> B) Lineal.
> C) Tetraédrica.
> D) Trigonal plana.
> E) Piramidal.

Esta pregunta inquiere sobre el conocimiento de un aspecto específico de la molécula de CS_2: La geometría de ésta. Sin embargo, si se considera el currículum de manera comprensiva, se tiene que una pregunta de este tipo es escasamente relevante en función de los objetivos de aprendizaje (aún cuando ésta, por alcance temático, pudiera formar parte del CMO de 2° medio "el enlace químico") [9]. Esto se debe a que los propósitos de aprendizaje asociados a este contenido no se refieren al dominio de conocimientos o datos específicos. Estos propósitos de aprendizaje están orientados a la capacidad de establecer cierto tipo de relaciones. Específicamente, si se consideran los OF, se tiene la inclusión de este contenido se relaciona con la capacidad de "[r]elacionar la estructura electrónica del átomo con su capacidad de interacción con otros átomos"[10]. No obstante, se puede apreciar que a través de esta pregunta no se está midiendo algún aspecto de la capacidad de relacionar señalada en este OF.

Pese a estas observaciones, se reconoce que la metodología utilizada supone un intento de responder a los OF al momento de elaborar la prueba. De acuerdo a la definición operativa de referencia curricular adoptada en los marcos evaluativos (y previamente citada en este documento), se establece que cada ítem de la prueba se relaciona tanto con un CMO como con una "acción pedagógica" informada a partir de los OF. Sin

[8] DEMRE (2006) Marco Teórico de la Prueba de Matemática (pág. 9).
[9] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 167).
[10] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 166)

embargo, como se verá en el punto siguiente, la forma en que se abordan estos OF no garantiza que sean efectivamente capturados y transferidos al instrumento de evaluación.

b) Análisis curricular de los OF

Para efectos de analizar las habilidades a ser evaluadas en la prueba, los marcos evaluativos analizan lo OF en términos de las "acciones pedagógicas"[11] que estarían contenidas en éstos. Estas acciones son extraídas individualizando el verbo que contiene cada OF, dejando de lado todo el resto de la formulación que constituye cada uno de estos objetivos. Posteriormente estos verbos son clasificados en distintas categorías de habilidades. Sobre la base de la distribución de estos verbos en las distintas habilidades, la metodología analiza la presencia que éstas poseen en el currículum nacional.

La focalización en los verbos que este procedimiento contempla resulta altamente problemática desde un punto de vista curricular. Al centrarse exclusivamente en este elemento en particular, la metodología de análisis utilizada *asume que es el verbo por sí solo el referente que permite dar cuenta de la habilidad que cada uno de estos OF supone*. Al operar de esta manera, se desconoce que los OF son formulaciones completas que no son reducibles a un verbo específico. Cuando cada verbo es tratado de manera aislada se pierde el sentido que cada aprendizaje tiene en el currículum, y se termina designando una acción planteada en términos abstractos, es decir, como una acción genérica que no logra describir el carácter del aprendizaje a lograr. Por otra parte, los OF *no son sólo una combinación entre una acción y un contenido*. Cada uno de ellos constituye una *descripción* de un aprendizaje a ser desarrollado. En tanto son una descripción de aprendizaje, su sentido se define a través de las condiciones y elementos incluidos en la formulación completa de éstos. En otras palabras, es la totalidad del OF la que describe el aprendizaje a desarrollar.

Como resultado, se tiene que los marcos evaluativos para la PSU, si bien toman como punto de partida los referentes centrales que el currículum contempla (los OF), terminan por introducir un procedimiento que invita a que tales referentes terminen siendo desvirtuados o desdibujados. Esta situación implica que *por medio de la focalización en ciertos verbos considerados de manera aislada se puede terminar señalando algo distinto a lo que originalmente plantea el currículum, o bien, favoreciendo la evaluación de aprendizajes que son irrelevantes de acuerdo a lo planteado en éste*.

A modo de ejemplo, en el caso de Lenguaje y Comunicación se señala como uno de los OF de 2° Medio "Interpretar el mundo creado en las obras, apreciando la diversidad de

[11] El uso del término "acciones pedagógicas" resulta confuso toda vez que una acción pedagógica consiste en un cierto procedimiento a ser desarrollado por el docente para lograr un cierto fin de aprendizaje. No obstante, los OF no refieren a aquello que se espera el docente realice. Estas son formulaciones que designan los objetivos de aprendizaje a los que se espera que los y las estudiantes logren llegar. Es decir, definen un fin, no un procedimiento o acción de carácter pedagógico para promover su logro.

mundos e interpretaciones posibles que ofrece la literatura". A partir de este OF, el marco evaluativo considera sólo el verbo "interpretar", y le confiere el carácter de una habilidad a ser evaluada, la que puede ser tratada con independencia del propósito de captar "el mundo creado en obras", como el OF especifica. Esto se puede observar en el siguiente ítem presentado en el documento "Marco Teórico Prueba de Lenguaje y Comunicación" (pág. 146):

Cuando el emisor dice "el sueño de la unidad prebabélica nunca ha cesado" se refiere a

A) una utopía del hombre que quiere comunicarse con Dios.
B) la aspiración de los pueblos por imponer su lengua.
C) la universalidad lograda por el latín como lengua oficial de la Iglesia Católica.
D) el anhelo del hombre de que exista un único idioma universal.
E) un sueño de la humanidad por comprender una lengua que capte toda la realidad.

En este ejemplo se puede apreciar que efectivamente los estudiantes deben llevar a cabo una interpretación. Sin embargo, se les pide interpretar el sentido de una expresión específica tomada del texto leído: "el sueño de la unidad prebabélica nunca ha cesado", y no una interpretación más global, que aluda a la interpretación de los mundos creados en las obras, que es lo que señala el OF anteriormente mencionado. De esta forma, se puede apreciar que el foco en el verbo o acción a partir del cual esta pregunta es elaborada no garantiza capturar la naturaleza de los aprendizajes a los que el currículum apunta.

c) Criterios de priorización y selección

Los marcos evaluativos suponen procedimientos claramente orientados a que la prueba finalmente logre reflejar de la manera más fiel posible el currículum. Esta intención se aprecia en la forma de acuerdo a la cual se definen los pesos relativos o el porcentaje de preguntas que se asignará a cada categoría de contenidos y de habilidades. Se puede apreciar que en los marcos evaluativos se busca intencionar que la prueba en cierta medida priorice estos elementos de acuerdo a la presencia o énfasis que éstos poseen en el currículum. Esta apreciación deriva al constatar el uso del análisis curricular presente en los mismos marcos de evaluación como un referente para este proceso. En éste se analiza el currículum en términos de las diferencias en el número de OF y de CMO en cada categoría de habilidad y contenido respectivamente.

Sin embargo, cabe levantar la pregunta respecto de la necesidad de considerar criterios complementarios a los que se derivan de este análisis cuantitativo. Específicamente, adquiere sentido plantear la conveniencia de tomar en cuenta criterios que se articulen en torno a las propiedades y características de los aprendizajes que son propios y específicos para cada sector de aprendizaje. Esto es, considerar para la asignación de pesos relativos la naturaleza de estos aprendizajes, y no sólo el conteo de habilidades y contenidos. Cabe mencionar que si bien los marcos evaluativos desarrollan también una

mirada cualitativa sobre el currículum, no se verifica luego que las categorías y consideraciones que se desprendan de dicho análisis se utilicen como referentes al momento de articular la prueba y sus componentes.

La conveniencia de otorgar una mayor consideración a la visión sobre la naturaleza de las habilidades propias de cada sector resulta importante por diversos motivos. En primer lugar, al realizar el análisis cuantitativo de los de los OF no se distinguen las *diferencias que puede haber entre éstos en términos de profundidad y extensión*. Al realizar la sumatoria, éstos son tratados como unidades equivalentes. Lo mismo sucede en lo que refiere al análisis de los CMO. Sin embargo, estos elementos no son necesariamente uniformes en términos de su envergadura y nivel de integración. Como resultado, estas categorías no constituyen unidades cuya relevancia y peso relativo se pueda derivar de una simple agregación de las mismas. A modo de ejemplo, para la prueba de Ciencias el CMO de Física "Pulsaciones entre dos tonos de frecuencia similar" no puede ser equiparable en términos de extensión y profundidad con el CMO "Distinción entre ondas longitudinales y transversales, ondas estacionarias y ondas viajeras"[12]. Esto implica, la asignación de un mismo valor a cada uno de ellos pasa por alto diferencias que requieren ser consideradas para efectos de la evaluación.

Por otra parte, es posible considerar que para efectos de la finalidad de la prueba no son todos los elementos del currículum igualmente relevantes. Es posible esperar que algunos de ellos resulten más significativos tanto para la función de selección universitaria, como para dar cuenta de aquello que se espera como logro final del conjunto de la Educación Media en cada sector. Por ejemplo, eventualmente podría argumentarse que *aquellos aprendizajes que se definen para los últimos años de la Enseñanza Media, debido a su mayor nivel de exigencia y de integración, resultan más apropiados para estos fines que aquellos correspondientes a los primeros años de este nivel educativo*. Esto último es un argumento que resulta especialmente pertinente en consideración del reciente ajuste curricular. Esto se debe a que dicho ajuste está organizado en con el propósito explícito de definir una progresión de aprendizajes creciente en términos de su complejidad e inclusividad.

Independiente de que esta sea la distinción más adecuada o no, el punto que se señala con estas observaciones apunta a la conveniencia de considerar la naturaleza o el carácter sustantivo de los aprendizajes propios del sector para definir criterios para seleccionar y priorizar qué se evaluará en la prueba.

[12] Ambos contenidos se registran en la tabla 2.1, página 15 del Marco teórico del subsector Física.

1.3 Sugerencias y propuestas

Las observaciones formuladas en este capítulo apuntan a la necesidad de considerar dos aspectos complementarios para efectos de fortalecer los marcos evaluativos y el alineamiento de la PSU con las bases curriculares. Por una parte, estas observaciones apuntan a la necesidad de considerar los OF como el elemento articulador central para establecer la referencia curricular de la prueba. En segundo lugar, se indica la conveniencia de considerar nuevos criterios para seleccionar y para asignar pesos relativos a aquello a ser evaluado, así como la forma en que el currículum es analizado con dicho fin.

Sin embargo, se reconoce que la transformación de estas observaciones en propuestas concretas supone un desafío y una dificultad importante dado que implican necesariamente, como es el caso de cualquier evaluación referida al currículum, un ejercicio de interpretación de las bases curriculares. Esta interpretación resulta necesaria para efectos de derivar categorías y tomar decisiones que permitan operacionalizar adecuadamente el currículum para los propósitos de evaluación.

Considerando que el desafío para efectos de evaluación supone el desarrollo de una mirada y análisis del currículum con el objeto de potenciar la sujeción curricular de la prueba, desde la UCE se ofrece la posibilidad de brindar apoyo y desarrollar líneas de trabajo conjunto en función de los siguientes aspectos:

a) En primer lugar, entablar una relación de cooperación para incorporar en los marcos evaluativos categorías y criterios en los que se logre expresar aquellos elementos que resultan más significativos desde un punto de vista curricular. Esto es, cooperación para complementar los marcos evaluativos de manera tal de reforzar la interpretación de currículum que se puede lograr a través de éstos.

b) En segundo, también con el fin de promover el alineamiento de la prueba con el currículum, entablar una línea de trabajo en función de los ítemes elaborados para la prueba. Específicamente, se la posibilidad de generar una línea de trabajo orientada a entregar retroalimentación y observaciones referidas a la forma en que éstos se relacionan con el currículum.

Parte II. Análisis Específico de Los Marcos Evaluativos

Esta parte del informe da a conocer el análisis llevado a cabo para cada uno de los marcos evaluativos de las cuatro pruebas. Estos análisis expresan la forma concreta en la que se manifiestan las observaciones arriba planteadas dentro cada uno de dichos marcos. Junto con esto, dan a conocer también aquellas observaciones de carácter específico que resultan pertinentes dadas las particularidades de cada uno de los sectores de aprendizaje evaluados. De tal forma, esta parte del documento se organiza en cuatro capítulos. El Capítulo 2 presenta el análisis realizado en torno al marco evaluativo para la prueba de Lenguaje y Comunicación; el Capítulo 3 presenta las observaciones realizados en el caso de Matemáticas; el Capítulo 4 el análisis para la prueba de Ciencias; y el Capítulo 5 para la de Historia y Ciencias Sociales.

Como se señaló en la introducción, la presentación de estos análisis no se organiza de acuerdo a una estructura común. Los distintos equipos a cargo de éstos utilizaron aquella que se consideró más pertinente para cada caso.

Capítulo 2. Análisis de marco evaluativo para prueba de Lenguaje y Comunicación

Introducción

El presente informe reporta el análisis elaborado desde la Unidad de Currículum y Evaluación al Marco de Evaluación de la prueba PSU de Lenguaje y Comunicación (PSU-L), instrumento de evaluación que condiciona el ingreso a las universidades del Consejo de Rectores y que es elaborada por el Departamento de Evaluación, Medición y Registro Educacional (DEMRE) de la Universidad de Chile.

Este análisis se articula en torno a dos puntos centrales. El primero de ellos consiste en el alineamiento del Marco de Evaluación con el currículum vigente. Es decir, este análisis considera el grado en que la propuesta elaborada por el DEMRE favorece la construcción de una prueba que efectivamente evalúe los aprendizajes que el currículum nacional define para el sector Lenguaje y Comunicación.

El otro punto considerado en el análisis refiere a las características técnicas del Marco de Evaluación. Este último es examinado en función de las propiedades que se esperaría encontrar en un documento de esta naturaleza.

El desarrollo de las observaciones se organiza en tres apartados, cada uno de ellos focalizado en un aspecto del Marco de Evaluación. Estas partes son:

1) **Metodología utilizada para el análisis del currículum**. En este apartado se tratan aspectos relevantes sobre las decisiones metodológicas adoptadas por el DEMRE para segmentar y cuantificar el currículum vigente como paso previo para la construcción de una matriz de evaluación vinculada con el currículum.

2) **Concepto de referencia curricular.** En este apartado se comentan aspectos importantes sobre la decisión metodológica del DEMRE para satisfacer el mandato de realizar una prueba con referencia curricular efectiva.

3) **Matriz curricular para la elaboración de ítemes**. En este apartado se comentan aspectos relativos a la conformación de la matriz de evaluación y las tablas de especificaciones para la elaboración de ítemes de la prueba PSU.

1. METODOLOGÍA DE ANÁLISIS CURRICULAR.

Los OF y CMO son ordenados en tres categorías de análisis: Lengua Castellana y Comunicación, Literatura, Medios masivos de comunicación. Estas categorías funcionan como ejes para separar y cuantificar la presencia de los OF y CMO de cada nivel. (Ver resumen de tablas I a IV en pp. 16 a 19).

Inicialmente el análisis aborda los OF y CMO de manera separada, los que finalmente son integrados en una tabla que busca establecer relaciones entre ellos y agruparlos de acuerdo a los tres núcleos temáticos arriba señalados (Ver tabla IX).

Al considerar este análisis en términos del grado en que permite capturar y reflejar el currículum de este sector surgen dos aspectos que deben ser tomados en cuenta:

1º Al establecer un análisis que separa los OF y CMO se abandona una perspectiva integradora de los mismos. Es importante tener en cuenta, en este sentido, que el currículum nacional intenciona los OF como ejes basales de los aprendizajes, en la medida que son "competencias o capacidades que los alumnos y alumnas deben lograr al finalizar los distintos niveles de la Educación Media y que constituyen el fin que orienta al conjunto del proceso de enseñanza-aprendizaje"[13]. Los CMO, por otro lado, operan como "conocimientos específicos y prácticas para lograr habilidades y actitudes que los establecimientos deben obligatoriamente enseñar, cultivar y promover para cumplir los objetivos fundamentales establecidos para cada nivel"[14]. Esto implica que los CMO requieren ser considerados en función del sentido que adquieren en relación a los OF. En el caso de la prueba de Lenguaje y Comunicación, específicamente, los CMO deben considerarse en función de la propuesta que subyace a los OF y que enmarca los aprendizajes que estos establecen. Dado esto, un análisis curricular en el que no se rescate esta relación puede traducirse en una lectura parcial del currículum. Uno de los riesgos que esta situación conlleva es el de interpretar los CMO como "contenido puro" a ser evaluado, situación que resulta siempre cuestionable al visualizar los CMO desde la perspectiva de los OF. Por ejemplo, uno de los CMO que el Marco de Evaluación señala como "Medible" y que se identifica con el código 102 (ver p.103) se formula de la siguiente forma: "Participación en situaciones privadas y públicas de interacción comunicativa, dando oportunidad para: El reconocimiento de relaciones de simetría y complementariedad entre los participantes; evaluación de las situaciones en que se dan tales relaciones que permita su modificación". Este CMO podría ser considerado desde el punto de vista de los conceptos lingüísticos evaluables que contiene, como **reconocer** "relaciones de simetría y complementariedad". Sin embargo, la pertinencia de esta interpretación no queda tan clara al considerar la perspectiva que promueve el OF: "Comprender los procesos de comunicación centrados principalmente en el intercambio de información y en la interacción entre pares". Este objetivo dirige el CMO hacia una habilidad de comprensión de la interacción, más que al reconocimiento de los conceptos lingüísticos implicados.

[13] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 7).
[14] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 8).

Pese a la separación entre OF y CMO, el análisis presenta también una forma de establecer una relación entre estos (Tabla IX). Sin embargo, si bien este mecanismo asegura cierta vinculación y correspondencia de los OF con los respectivos CMO de cada nivel, no es del todo adecuada la manera en que la relación entre ambos elementos es planteada. Esto debido a que esta relación se formula en términos de una correspondencia lineal entre cada CMO con un determinado OF. Es decir, se asume mediante esta vinculación que cada CMO es la especificación de un determinado OF, dejando de lado posibles vínculos o relaciones que puede haber con otros OF que el currículum define. El sentido de los CMO no necesariamente se agota en su relación con un OF específico.

2º Otra observación relativa a la metodología de análisis curricular tiene que ver con la distribución por categorías adoptado. Como se señaló anteriormente, los ejes que este marco contempla son: a) Lengua Castellana y Comunicación; b) Literatura; c) Medios Masivos de Comunicación. A partir de estos ejes se agrupan los OF y luego se agrupan en cada OF los CMO correspondientes de cada nivel. Llama la atención que se adopte esta decisión de análisis, siendo que el currículum nacional ya tiene una separación en cuatro partes para los CMO. Específicamente, la diferencia se genera en relación al primero de los ejes arriba señalados: Lengua Castellana y Comunicación[15]. En lugar de este eje, el currículum define dos tareas de desarrollo distintas: Comunicación Oral y Comunicación Escrita. De esta forma, el Marco de Evaluación funde las dos tareas de desarrollo recién señaladas (Comunicación Oral y Comunicación Escrita) en uno solo eje (Lengua Castellana y Comunicación), lo que origina una lectura de los CMO demasiado acotada a contenidos "medibles" y que, por lo mismo, escapan a los sentidos que estos adquieren en relación a los OF.

Por último, es necesario hacer un alcance sobre los conceptos en torno a los cuales los OF son analizados, específicamente, sobre el uso del concepto de "acciones pedagógicas" en dicho proceso. Resulta confuso el uso de este término de manera equiparable al de habilidades (ver pág. 26). ¿Cómo se justifica esta opción? Se entiende que con este concepto el marco evaluativo pretende agrupar principalmente habilidades y actitudes, como *comprender/valorar,* (ver tabla p. 67), pero este término confunde porque "acciones pedagógicas" se asocia directamente con la descripción de lo que el profesor debe hacer y no con lo que el estudiante debe aprender.

2. NOCIÓN DE REFERENCIA CURRICULAR.

Sobre la noción de referencia curricular, el Marco de evaluación de la PSU señala que esta se establece el año 2002 por la "Mesa Escolar, entidad que se articuló como un mecanismo relacional entre el DEMRE, el Mineduc y el Comité Técnico Asesor del Honorable Consejo de Rectores" (p. 84). La perspectiva sobre la que se levantó este concepto apunta a la necesidad de considerar que las pruebas de selección no pueden enfocarse solo desde la educación superior, puesto que este criterio va en detrimento de la educación superior misma y también de la educación secundaria. Por tanto, se

[15] Las otras dos tareas de desarrollo definidas en el marco curricular son coincidentes con los ejes considerados en el Marco de Evaluación. Éstos son Literatura y Medios masivos de Comunicación.

plantea la necesidad de considerar a la prueba de selección universitaria como "un dispositivo de conexión entre educación escolar y superior, con funciones de selección y de evaluación curricular del sistema productor de las capacidades en base a las cuales se realiza la selección" (p. 84). La noción de referencia curricular surge así como necesidad de esta exigencia que el sistema de selección se impone. En consecuencia, la manera como se materializa esta relación implica que "cada uno de los ítemes de la prueba está basado en uno de los CMO, así como en una habilidad cognitiva" (p. 85). Por otra parte, se define que las habilidades cognitivas "se han recogido, en parte del constructo teórico anterior de la PAA-V, así como de las *acciones pedagógicas* que informan los Objetivos Fundamentales del subsector" (p. 85). Se plantea que este concepto puede tener dos acepciones:

Referencia curricular temática: aquella que aborda los CMO solo como categorías semánticas, privilegiando la medición de las habilidades cognitivas (PAA tradicional).
Referencia curricular efectiva: aquella que se centra en los CMO, midiéndolos desde las habilidades cognitivas (la actual PSU en su fase de transición).

La noción de referencia curricular así entendida supone un desafío. El objetivo es reconocer los CMO del Marco y considerarlos en la PSU desde un conjunto de habilidades (13 en este caso) que ha sido identificado. De este concepto resulta otro que es fundamental para definir el carácter de la PSU. Por tratarse de un instrumento de selección que considera un CMO desde alguna habilidad cognitiva, esta prueba se define como prueba de razonamiento, en tanto combina estos dos componentes. El hecho de que se trate de una prueba de razonamiento supone que se "evalúan las habilidades cognitivas y los modos de operación y métodos generales aplicados a la resolución de problemas" (p. 92). Al aplicarlo de este modo, el concepto de referencia curricular delimita el carácter de la PSU: una prueba que debe considerar una serie de habilidades, desde las más simples a las más complejas, para resolver problemas propios de la naturaleza de la disciplina.

Junto con ello, esta referencia se busca lograr también por medio de la asociación de los ítemes de la prueba con los Programas de estudio elaborados por el Mineduc (ver pág. 86). La justificación presentada para utilizar estos programas como referentes para la construcción de ítemes alude a que el documento base del currículum no considera las variables didácticas para la concreción de los OF-CMO, mientras que los programas de estudio sí ofrecen orientaciones de este tipo.

Esta forma de concretizar la noción de referencia curricular supone necesariamente un acercamiento de la prueba al currículum. Sin embargo, emergen ciertas dudas respecto del grado en que efectivamente es posible lograr un alineamiento de la prueba con el currículum nacional sobre la base de esta aproximación. Esto, debido a que esta operacionalización de la referencia curricular no captura o integra algunos elementos clave del currículum, y a que junto con ello, favorece la incorporación de elementos que son ajenos o no completamente ajustados al mismo. Esta situación se verifica en relación a dos aspectos de esta opción de referencia curricular.

La primera de ellas guarda relación con la ausencia de referencia a los OF, y a la inclusión de habilidades en lugar de estos. En relación a este punto, es necesario recordar que el marco evaluativo presenta la referencia curricular en términos de la relación de cada ítem con un ***CMO*** y con una ***habilidad***; visión que deja de lado una

alusión y consideración directa de los OF. De esta forma, el concepto de referencia curricular adoptado diluye la posibilidad de elaborar ítemes que se sustenten sobre los principales referentes del currículum del subsector, situación que puede ocasionar una visión parcial o no ajustada de las orientaciones del mismo.

Esta observación no pasa por alto el hecho de que el marco evaluativo propone cierta forma de integrar los OF en la operacionalización de la referencia curricular. Esto se aprecia en la inclusión de las habilidades que se lleva a cabo. Estas habilidades, como se señala en la página 99, derivan del análisis que el documento presenta en su primera parte. De acuerdo a dicho análisis, las habilidades son formuladas a partir de los OF, específicamente, estas son definidas en términos de los verbos que se incluyen en la formulación de estos objetivos. Junto con ello, se señala que estas habilidades están asociadas también a la taxonomía de Bloom.

No obstante, la forma en que se conciben e integran estas habilidades supone una doble dificultad para dar cabida a la evaluación de los OF en la prueba. Por una parte, estas limitaciones se constatan si se sigue la misma lógica contemplada para el análisis curricular, caracterizada por concebir la habilidad contenida en el OF en función del verbo utilizado para la formulación del mismo. Esto porque se observa finalmente que la categorización de habilidades a la que se llega no incluye la totalidad de estos verbos.

Habilidades cognitivas	Verbos que encabezan los OF en el currículum vigente
Conocer Comprender-analizar **Identificar** Caracterizar Analizar-sintetizar Analizar-interpretar Inferir localmente Sintetizar localmente Sintetizar globalmente **Interpretar** Inferir globalmente Transformar Evaluar	Afianzar Alcanzar Analizar Apreciar Comprender **Conocer** Crear Descubrir Desempeñarse Elaborar Explorar Expresar Fomentar Formar Fortalecer **Identificar** Incrementar **Interpretar** Investigar Producir Proponer Reconocer Reflexionar Relacionar Reparar Tomar conciencia Utilizar Valorar

Como se ve, de los 27 verbos utilizados en los OFV, solo tres (conocer, identificar e interpretar) coinciden totalmente con las habilidades cognitivas de la PSU. Tres de estas (comprender-analizar, analizar-sintetizar, analizar-interpretar) coinciden parcialmente.

Sin embargo, existe una diferencia más profunda y crítica con el currículum. La formulación de habilidades en términos de un verbo expresado en su infinitivo no refleja necesariamente la habilidad y el aprendizaje hacia el cual los OF apuntan. Este es el caso incluso en aquellas ocasiones en que la habilidad es formulada en función del verbo explícitamente señalado en un OF. Por ejemplo: En la pregunta que muestra la capacidad de interpretar (Marco Teórico PSU, p.146) se pide interpretar el sentido de una expresión. En la terminología utilizada en la PSU, debió haberse inventado una nueva habilidad, como es "interpretar localmente". En el currículum vigente, en cambio, en el OF 8 de 2° Medio se lee: "Interpretar el mundo creado en las obras, apreciando la diversidad de mundos e interpretaciones posibles que ofrece la literatura". Como se ve, ambos documentos dirigen la interpretación a direcciones distintas.

Adicionalmente, se observa poca claridad en los sentidos específicos otorgados a algunas habilidades, así como a las diferencias entre ellas. Este es el caso de la distinción que se hace entre conocer (código 001 de Tabla de especificación de página 99) e identificar (código 003 de Tabla de especificación de página 99), pues no es explícita la diferencia entre "saber información explícita del texto" y "reconocer elementos (…) presentes en el estímulo." En ambos casos se requiere que el estudiante realice un proceso de localización de información explícita en el texto. También resulta poco clara la distinción que se hace en la página 99 en los descriptores Analizar-sintetizar (código 005) y Sintetizar globalmente (código 009). Lo mismo ocurre con la distinción analizar- interpretar (código 006) con interpretar (código 010). Por otro lado, la habilidad cognitiva que aparece en los descriptores está simplemente nombrada, sin mayor especificación o explicación. Esto la hace vaga. En el ejemplo propuesto en la página 89 se tiene como descriptor "Conocer", sin especificar de qué tipo de conocimiento se trata (de hechos, de terminología, de principios, etc.). Una mayor precisión en el sustento teórico ayudaría a una mejor selección y variedad de las preguntas.

Aparte de las observaciones arriba señaladas en torno a la dilución de los OF y a la inclusión de habilidades, existen también observaciones relativas al uso de los programas de estudio para la operacionalización de la referencia curricular de la prueba. Si bien estos programas constituyen una opción realista para desarrollar una bajada pedagógica efectiva del currículum, su uso para efectos de la construcción de la prueba redundaría en trabajar los ítemes de la PSU a partir de una interpretación del currículum y no a partir de la fuente primera que es el marco en sí mismo. Esto, debido a que los programas constituyen siempre una lectura -válida y coherente, por cierto- del referente curricular de base con el objeto de hacerlo operativo, pero no son equiparables con este. En consecuencia, el uso de estos programas puede introducir cierta desviación en relación al marco. Por último, la interpretación de los programas, si bien es un derivado adecuado y validado de este último, no corresponde necesariamente a la lectura realizada por todos los establecimientos, ya que hay algunos que trabajan con programas de elaboración propia, lo cual es tan valido y coherente como los programas del Mineduc.

3. MATRIZ CURRICULAR PARA LA ELABORACIÓN DE ÍTEMES.

En este capítulo se analizan aspectos referidos a la matriz curricular presentada para transferir el análisis curricular previamente desarrollado a la elaboración de la prueba propiamente tal. A diferencia de los apartados anteriores, este punto es analizado no solo en función del alineamiento curricular, sino también en función de las características técnicas del mismo. El análisis de estos aspectos técnicos, no obstante, no debe entenderse como un elemento separado o intrascendente desde el punto de vista del alineamiento con el currículum. Por el contrario, como se presentará a continuación, existen algunas implicancias importantes en términos de esto último.

3.1.- Sobre el concepto de "Matriz curricular" y "Tabla de especificaciones"

En el capítulo V, denominado "Matriz curricular de Lengua Castellana y Comunicación para la elaboración de ítemes en la PSU de Lenguaje: tabla de especificaciones", se hace referencia a "Tabla de especificaciones" y "Matriz curricular" de manera indistinta o al menos imprecisa, pues en la página 94 se afirma:

"La Tabla de especificaciones para la prueba de Lenguaje que aparece en este apartado es la sistematización lograda luego de la caracterización y el análisis del Marco Curricular de Lenguaje y Comunicación y la noción de referencia curricular que se expone al final de este documento.

Tal como hemos visto, a fin de poder medir los CMO desde las habilidades cognitivas, se hace necesario construir una matriz curricular, entendiendo que ésta constituye – mínimamente – una tabla de doble entrada, desde la que surgen los espacios o casilleros en los cuales se construirán los ítemes que darán forma a los distintos instrumentos de evaluación."

En el primer párrafo citado se presenta la fuente de la Tabla de especificaciones (la sistematización y análisis del currículum). De esto se puede entender que las tablas de especificaciones son las que se encuentran entre las páginas 98 y 137, las cuales no especifican lo evaluado, sino que desglosan los contenidos del currículum.

En el segundo párrafo citado se describe la forma y organización de la matriz curricular. De esta descripción se puede entender que la matriz curricular es la tabla que se encuentra en la página 139, sin embargo el título de dicha tabla es "Tabla de especificaciones de la prueba de Lenguaje y Comunicación".

Por lo anterior, se hace necesario precisar el marco conceptual, pues, por ejemplo en el contexto de la prueba SIMCE, la matriz curricular es una tabla de porcentajes de las habilidades o procesos cognitivos evaluados; y la tabla de especificaciones describe los

objetivos de evaluación de cada ítem de una prueba. En otras palabras, la matriz curricular es un documento general sobre las características de la prueba y la tabla de especificaciones, como su nombre lo indica, es una especificación del objetivo de evaluación de cada ítem o grupo de ítemes donde se explicita qué es lo que se evalúa, qué se espera de los estudiantes que responden a determinada pregunta o preguntas. Podría resultar útil homogenizar estos conceptos entre los distintos sistemas de evaluación con referencia curricular.

3.2.- Sobre objetivos de evaluación y la expectativa curricular.

El objetivo de las tablas de especificaciones en un marco de evaluación debería ser mostrar claramente cómo las habilidades asociadas con los conocimientos de la disciplina se combinan para hacer un constructo coherente, con el fin último de dar a conocer qué es lo que se evalúa. En pruebas internacionales, referidas al curriculum, por ejemplo A-Level (Inglaterra)[16], se presentan especificaciones claras a partir de lo establecido en el currículum.

Por otro lado, en la página 165 del Marco de evaluación PSU se explicita:

"…los instrumentos de evaluación requieren de un escenario evaluativo propio, con indicaciones precisas de qué, cuánto y cómo evaluar, pero en el entendido de que dicho escenario mantiene su foco en el mismo centro que el Marco Curricular."

El problema conceptual referido al concepto de matriz curricular y tabla de especificaciones hace que el capítulo referido a la matriz curricular no muestren una clara alineación al espíritu u objetivo global expuesto en el currículum, pues las tablas que se presentan no dejan claro su objetivo: ¿precisar lo que se evaluará?, ¿o desglosar el currículum para mostrar el proceso de elaboración? Por otra parte, no logran reflejar claramente los aprendizajes que se espera evaluar. Esta opinión se fundamenta al observar los cuatro tipos de tablas que se presentan en el capítulo.

El primer tipo de tabla, en la página 98, presenta una matriz general donde se explicitan los tres grandes focos de evaluación de la prueba y la cantidad de ítemes asociados a cada foco. Si bien es cierto, esta tabla constituye una matriz curricular donde se observan las características generales de la prueba, no explicita el objetivo de cada sección, ni la referencia curricular. En la Sección 1 se podría explicitar cada eje evaluado y su objetivo de evaluación y en las secciones 2 y 3 los objetivos globales. Todo esto para describir de manera general los objetivos y relacionar la prueba con el currículum nacional. Por otro lado, no se explicita que esa es la matriz curricular, nosotros debemos inferirlo, se sugiere poner el título.

El segundo tipo de tabla que se incluye en el capítulo, en la página 99, presenta especificaciones para la construcción de pruebas referidas a habilidades y competencias cognitivas del Curriculum de Lengua castellana y Comunicación. Esta tabla contiene todas las habilidades planteadas en el currículum.

[16] Sitio www.aqa.org.uk

Al observar esta tabla de especificación referida a las habilidades del currículum se observa el siguiente problema: las habilidades se presentan aisladas de los OF del currículum, por lo tanto resulta ser una tabla abstracta que no demuestra qué se evaluará específicamente en los ítemes de la PSU. Se espera que una tabla de especificaciones muestre las habilidades y OF relacionados coherentemente y que se visualicen los ítemes de la prueba a través de la declaración explícita del objetivo de evaluación del ítem.

El tercer tipo de tabla que se incluye en el capítulo, desde la página 101 a la página 137, presenta especificaciones referidas al componente de contenidos del currículum para cada nivel. Al observar estas tablas de especificaciones referidas a contenidos se observa el siguiente problema:

En las tablas de especificaciones se desglosan los 86 CMO del currículum y se agregan los contenidos planteados en los programas. Si bien se explicita con la sigla NM qué contenidos no son evaluables o medibles en el formato de la prueba, se afirma que las habilidades asociadas a estos contenidos sí pueden ser consideradas en la matriz curricular que configura la prueba de lenguaje (ver nota a pie de página número 16, página 103). Ante esto, nuevamente surge la interrogante sobre el objetivo de esta tabla de especificaciones, ¿precisar lo que se evaluará? o ¿desglosar el currículum para mostrar el proceso de elaboración? Además esta fragmentación no permite visualizar qué se espera que un estudiante pueda hacer con sus conocimientos y habilidades referidas al lenguaje.

El cuarto tipo de tabla, en la página 139, presenta un modelo de especificaciones de la prueba donde se deben asociar a las trece habilidades, los códigos pertenecientes a los CMO especificados en las tablas anteriores. Se observan dos problemas:

En primer lugar, los códigos no están presentes, por lo tanto no se explicitan los cruces entre CMO y habilidad, por esto la tabla de especificaciones solo es un 'modelo' y no una tabla de especificaciones propiamente tal.

En segundo lugar, en la nota 19 de la página 139 se afirma que existen 86 contenidos mínimos establecidos en el currículum a lo largo de la EM, de los cuales 60 son evaluables en el formato PSU. Ante este análisis cuantitativo surgen las preguntas:

¿Cuáles son los contenidos evaluables?
Si son todos los contenidos relevantes en el currículum, ¿qué criterio predomina? ¿Cómo se seleccionan los contenidos? En el caso de que todos los contenidos tengan igual relevancia, sería necesario explicitarlo mayormente.

Por otro lado, en la misma nota se afirma que esos 60 CMO se multiplican con las 13 habilidades cognitivas planteadas en el currículum. Ante esto, surge otra pregunta:
¿Todas las combinaciones son posibles y relevantes?

El problema fundamental de esta separación de habilidades y CMO que solo se relacionan a través de un proceso de codificación en la tabla de la página 139, es que da a entender que todas las habilidades pueden ser asociadas a un CMO, lo que por lo tanto indica que habría 780 posibilidades de objetivos de evaluación. Al dejar abierta todas las posibilidades no se captura el sentido sustantivo de los aprendizajes que

deberían ser evaluados y se demuestra que el análisis del currículum se ha hecho desde una perspectiva fragmentaria del Lenguaje y Comunicación sin captar el sentido último explicitado en el currículum: que los estudiantes de enseñanza media puedan analizar, evaluar y construir significado en diferentes situaciones comunicativas.

La Matriz Curricular o Tabla de especificaciones referente a los contenidos no es concreta, nunca se puede con certeza saber cuál CMO es el que se releva o más bien qué se espera que sepa hacer el estudiante. Si bien se precisa que los contenidos seleccionados surgen de las aplicaciones experimentales que permitían ir delimitando los contenidos más conocidos por los estudiantes, dada la heterogeneidad de los establecimientos educacionales de Chile, sería necesario establecer criterios de selección referidos al currículum nacional y explicitarlo en este documento.

Por otro lado, el criterio de selección de un contenido determinado no debería enfatizar tanto la relevancia que le dan los profesores en la sala de clases, sino asignar un lugar mucho más importante a la relevancia que plantea el currículum.

Para que la tabla de especificaciones sea tal, debería redactarse asociando las habilidades a determinados Objetivos Fundamentales. Esto se puede hacer de manera general describiendo lo que se espera que haga un estudiante con lo aprendido durante los cuatro años o con lo que se espera que haga un estudiante en cada nivel, para eso son muy útiles las tablas presentadas en las páginas 54, 55, 56 del Marco de Evaluación donde se relacionan los OF y se puede observar la progresión de aprendizajes:

Algunas ideas de objetivos generales de evaluación:

- Analizar, evaluar los procesos de comunicación centrados en el intercambio de información, la exposición de ideas o en la controversia generada por la diferencia de opinión.(Tabla VI. 1 Secuencia de OF cada núcleo temático por Nivel, eje Lengua Castellana y Comunicación).

- Reconocer y evaluar elementos paraverbales y no verbales que se emplean en la interacción informativa, la exposición de ideas o argumentos. (Tabla VI .1 Secuencia de OF cada núcleo temático por Nivel, eje Lengua Castellana y Comunicación).

- Interpretar visiones de mundo expuestas en diversas obras literarias y relacionarla con el contexto de producción o recepción. (Tabla VI. 2 Secuencia de OF cada núcleo temático por Nivel, eje Literatura).

Las tablas donde se presentan las secuencias de OF de cada núcleo temático permiten visualizar el objetivo global del currículum, puesto que se explicita la progresión de las habilidades, temas y conocimientos que debe adquirir un estudiante de Enseñanza media.

En síntesis, debido a la imprecisión en los conceptos de “matriz curricular” y “tabla de especificaciones” en este marco de evaluación, el capítulo no deja claro los objetivos de evaluación. Por lo anterior, es recomendable, en primer lugar, precisar u homogenizar los conceptos.

No se observa una matriz curricular donde se vea la prueba completa con las habilidades y OF evaluados ni una tabla de especificaciones que describa el objetivo de evaluación de cada ítem, por lo tanto el lector de este marco no visualiza las características de esta prueba.

Con las tablas presentadas entre las páginas 98 y 139 se puede ver el proceso de organización de las habilidades, pero no se ve la meta final: una tabla con los objetivos de evaluación y habilidades relevadas.

El problema central del capítulo es que no se percibe una adecuada alineación al currículum, pues al centrarse en CMO y presentar una perspectiva fragmentaria del análisis curricular se percibe una interpretación sesgada de los planteamientos del Currículum de Lengua Castellana y Comunicación al no rescatar el espíritu o los objetivos generales de los estudios del Lenguaje.

3.3. Sobre la matriz curricular centrada en la sección 1 de la PSU

En el capítulo V del Marco de evaluación PSU no se visualiza una matriz que especifique las habilidades evaluadas en las secciones: Indicadores de producción de textos (Sección 2 de la prueba), Comprensión de lectura y vocabulario contextual (Sección 3 de la prueba), lo que es de suma importancia para determinar las habilidades específicas que se evalúan en toda la prueba. Por lo anterior, la sección 1 de la PSU, referida a contenidos correspondientes a Lenguaje y Comunicación, parece ser la parte más relevante de la prueba debido a que gran parte del marco se centra en eso. Esto es contradictorio con el objetivo del currículum centrado en habilidades y los porcentajes asignados a cada parte de la prueba. Es necesario explicitar una matriz curricular y una tabla de especificaciones para las secciones de "Indicadores de producción de textos" y "Comprensión de lectura y vocabulario contextual".

3.4. Sobre la focalización en contenidos en PSU

La PSU-L aparece presentada como una prueba de razonamiento, que a diferencia de la PAA, tiene 15 ítemes de conocimientos y habilidades básicas de Lengua Castellana y Comunicación. Los 65 ítemes restantes sobre conectores, plan de redacción, vocabulario contextual y comprensión de lectura, mantendrían la orientación de la PSU, como prueba de habilidades centrada en las habilidades cognitivas. Sin embargo, la PSU postula que las preguntas y los textos deberán establecer una relación significativa con alguno de los CMO de la matriz curricular. Este intento de vinculación lleva a seleccionar textos relacionados con los medios de comunicación (pp. 97-144) o a buscar en los textos ciertos contenidos de lo que los autores llaman Lengua Castellana (tipo de situación comunicativa, tipos de discurso, ortografía, tipos de texto). Esta búsqueda de vinculación hace que la segunda sección de la Prueba, Indicadores de la producción de textos, continúe con la comprobación de conocimientos específicos.

Así, en la "forma de la prueba utilizada en el Proceso de Selección a la Educación Superior 2006" entregada por la Universidad de Chile, en la sección Indicadores de producción de textos (10 preguntas), encontramos:
Dos ítemes sobre textos de los medios de comunicación.
Dos ítemes sobre teoría literaria.
Tres ítemes sobre el discurso y tipos de texto.
Un ítem sobre ortografía.
Dos ítemes sobre temas científicos.

Lo anterior muestra que 8 de las 10 preguntas de dicha sección se relacionan directamente con contenidos específicos de los CMO y los programas.

En la tercera sección: Comprensión de lectura (50 preguntas), se sigue la misma tendencia. De hecho, en la forma presentada, las siguientes preguntas, más que por la comprensión del texto, inquieren sobre conocimientos específicos:
Ítemes 44, 46 (teoría del discurso)
Ítemes 45, 64 (uso de los signos de puntuación)
Ítem 46 (teoría del discurso)
Ítems 47 (tipos de expresiones: literal, figurada, gramatical)
Texto 10 (54 –60) preguntas sobre características de los mensajes de los medios de comunicación.

En la prueba del 2003, en la segunda sección "Indicadores de producción de textos, Manejo de los conectores", las cinco preguntas (16-20) versan sobre temas literarios o gramaticales. En plan de redacción (21-30) ocho preguntas versan sobre temas literarios.

En la tercera sección "Comprensión de Lectura, Textos breves" (31-53):
El ítem 31 versa sobre el conocimiento de los tipos de mundo.
El ítem 36 requiere de conocimientos para responder sobre el tópico del viaje.
El ítem 42 requiere conocimientos sobre los registros de habla.
El ítem 43 requiere de algunos conocimientos sobre el realismo inglés y las técnicas narrativas de Proust y Joyce.
El ítem 45 pregunta por los tipos de viaje.
El ítem 46 exige conocer el concepto de monólogo interior.
El ítem 49 pregunta por la finalidad comunicativa.
El ítem 50 implica conocer el concepto de ciencia ficción.
El ítem 51 versa sobre los medios de comunicación.

En la tercera sección, "Textos extensos y vocabulario" (54-80) dos de los tres textos seleccionados, abordan temas lingüísticos y el tercero las características de los medios de comunicación.

Todo lo señalado indica que la relación de la PSU con los conocimientos específicos del Marco, va mucho más allá de las 15 preguntas específicas sobre ellos y se hace presente en toda la prueba. Esto la hace alejarse del objetivo general del currículum centrado en las habilidades comunicativas y lo lleva a centrarse en contenidos.

La inclusión de temas lingüísticos, literarios y sobre los medios de comunicación en los textos presentados en las secciones de "Indicadores de producción de textos y

Comprensión de Lectura y vocabulario contextual", plantea objetivos de evaluación distintos a la evaluación de habilidades, pues el estudiante debe integrar conocimientos conceptuales además de habilidades. Esto, desde la perspectiva de evaluación, no es adecuado, pues un ítem debe evaluar lo que se explicita en él. Si el objetivo del ítem es que el estudiante organice adecuadamente las ideas de un texto, no debería agregar dificultad a partir de los contenidos, a menos que lo explicite.

CONCLUSIÓN

Del análisis del currículum se desprende que el objetivo general de Lenguaje y Comunicación es que los estudiantes de EM comprendan, produzcan y participen en situaciones comunicativas óptimas, pues:

"(…) en la Educación Media el proceso de enseñanza- aprendizaje lingüístico-gramatical y ortográfico reiterará muchos de los contenidos de la Educación Básica e incorporará otros nuevos, conforme a las necesidades de comprender y producir discursos orales y escritos más complejos, significativamente contextualizados por los estudiantes." (Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media, 1998:38).

El Marco de evaluación de PSU-L debería tener esto como primera premisa (o explicitarlo en este documento) para guiar la elaboración de ítemes para la sección 1 "Conocimientos de conceptos básicos y habilidades generales de Lenguaje y Comunicación", que concentra gran parte de la atención de este marco teórico, de manera que los contenidos y habilidades evaluadas tengan una relación más clara con este objetivo general y principal que define el currículum para esta disciplina.

El Marco de Evaluación del DEMRE manifiesta un intento por centrarse en el currículum de Enseñanza media, pero no realiza una lectura que capture el sentido o el espíritu del currículum (OF), pues justamente por centrarse en contenidos conceptuales específicos, deja de lado el objetivo central que el currículum asigna a la disciplina, que es lograr que los estudiantes terminen la enseñanza secundaria siendo capaces de participar adecuadamente en situaciones comunicativas complejas.

Se hace necesario agregar en el capítulo VII precisiones conceptuales, pues en dicho capítulo se realiza una justificación teórica de los contenidos curriculares. En el contexto de un Marco de evaluación es esperable que se presenten o se expliciten lineamientos teóricos sobre el concepto de evaluación, habilidades y todo el aparataje conceptual sobre el que se sustenta la prueba.

Capítulo 3. Análisis de marco evaluativo para prueba de Matemática

Introducción

Este documento expresa tanto las reacciones al documento "Marco Teórico de la PSU Matemática" como las concepciones que los equipos de currículum y evaluación tienen con relación a los temas en diálogo. El énfasis está situado principalmente en responder las preguntas que el equipo PSU realiza, entendiendo que éstas constituyen un insumo de importancia para el análisis y mejoramiento del sistema de selección universitaria.

Respuesta a las preguntas planteadas

Los siguientes puntos explicitan cada una de las preguntas que el equipo PSU plantea y para las cuales los equipos de matemática, tanto de SIMCE como de Currículo, consideran son de su competencia. De esta forma el formato se ajusta al tipo, pregunta-respuesta, esperamos facilite la lectura y comprensión.

1. Observaciones respecto de la Metodología de Análisis del currículum.

La metodología utilizada en la organización de los OF/CMO con la finalidad de visualizar las categorías descritas en la página 14, permite hacer solamente un análisis cuantitativo de la primera categoría mencionada. Lo anterior debido a que el procedimiento principal para establecer tanto la secuencia como la relación entre OF/CMO y habilidades, categorías mediante las cuales se analiza la propuesta del sector de Matemática para la Enseñanza Media; distribuye los OF/CMO y las horas asignadas al estudio de los ejes temáticos, posteriormente se cuantifica la información que proporcionan esas distribuciones en una serie de tablas que organizan la información contenida en el currículum nacional reflejándose posteriormente en el Marco Teórico.

La cuantificación resultante conduce a un análisis incorrecto de la importancia y el sentido curricular de cada uno de los ejes identificados, dado que la cantidad de OF no determina necesariamente la relevancia. El formato de redacción actual de Objetivos Fundamentales da cuenta de grandes aprendizajes que se esperan lograr en tiempos que varían dependiendo del nivel de complejidad de dichos aprendizajes.

A modo de ejemplo, analizaremos el siguiente Objetivo Fundamental de Segundo Medio:

"Conocer y utilizar conceptos matemáticos asociados al estudio de la ecuación de la recta, sistemas de ecuaciones lineales, semejanza de figuras planas y nociones de Probabilidad; iniciándose en el reconocimiento y aplicación de modelos matemáticos".

Una primera mirada a éste nos muestra una visión integrada de saberes matemáticos relacionados con el álgebra (Ecuaciones de recta, Sistemas de ecuaciones) la geometría (semejanza de figuras planas, ecuaciones de recta) y probabilidades en función de promover habilidades relacionadas con el modelamiento matemático (reconocimiento y aplicación). Esta visión integradora de conocimientos da un marco interpretativo que supera a las categorías taxonómicas usadas en el informe en el marco evaluativo para la PSU-M. Es así, como el "conocer" expresado en este OF hace referencias a "un conocer en relación con" , conocimiento que cobra significado en ciertos contextos problematizados los que son modelados (representados), por ejemplo, mediante ecuaciones lineales, no es un "conocer" atomizado, desprovisto de contenido.

Por su parte, el marco teórico en referencia, en su intención cuantificadora, desagrega este objetivo y clasifica cada una de sus partes en uno de los ejes temáticos, esto se explicita en la tabla I.2 de la página 17.

Una situación similar se observa en "La tabla las acciones pedagógicas referidas en los OF por nivel y eje temático" (pág. 48), la cual permite solamente cuantificar la información contenida en ella y entregar las habilidades correspondientes a cada eje temático. Lo anterior no considera el hecho que las habilidades por sí solas no determinan grados de dificultad claramente definidos. El análisis realizado en el ejemplo anterior permite determinar diferencias significativas entre aprendizajes asociados a una misma habilidad.

De forma similar, habilidades jerarquizadas como de orden superior por las taxonomías usadas a mediados del siglo pasado, pueden mostrar niveles de complejidad inferior cuando se analizan los desempeños asociados. Por ejemplo, aplicar el algoritmo de la adición en ejercicios rutinarios resulta de menor complejidad que comprender que la matemática es un lenguaje que permite modelar situaciones del mundo social.

La metodología utilizada, que permite realizar sólo algunas cuantificaciones, proporciona un insumo insuficiente para la comprensión del currículo y sus alcances, específicamente en lo relativo a la interpretación de los Objetivos Fundamentales y sus reales niveles de complejidad.

2. Observaciones respecto a la Noción de referencia curricular para la PSU – MAT

Respecto de las habilidades cognitivas, es cuestionable el pensar que las dimensiones evaluadas en la anterior PAA representen en la actualidad una manera adecuada y técnicamente válida de monitorear el logro de los objetivos descritos en el currículum nacional vigente. Las habilidades cognitivas recogidas en parte del constructo teórico de la anterior PAA y PCE (pág. 56), tienen relación con un enfoque curricular derogado por ley hace más de diez años. El enfoque vigente sitúa los aprendizajes matemáticos más allá de una sola disciplina. Un cambio importante con relación al currículo pre-reforma es lo relativo a los contextos; mientras el currículo pasado ofrecía una matemática por la matemática, el marco vigente tensiona la resolución de diversas

situaciones problemáticas en contextos de la vida cotidiana. Esta diferencia, en el proceso de enseñanza-aprendizaje, se constituye en un elemento de discontinuidad importante entre los aprendizajes promovidos por el currículo en la escuela y el tipo de evaluación que genera la selección a las Universidades tradicionales. Por lo anterior, no es extraño que el sistema destine una importante cantidad de recursos a la preparación de la prueba de selección, generando, en los últimos años de escolaridad, un currículo paralelo. [17]

3. Relación de la PSU – MAT con el Currículum.

La cuantificación mostrada en el documento "Marco Teórico PSU" muestra que la prueba en sí, puede medir el 33% de las combinaciones entre CMO y habilidades cognitivas del referente curricular, una vez descontados los 9 contenidos que aparecen en el currículum que se consideran no medibles por evaluaciones de ítems cerrados.
En relación con el 33% de las combinaciones medibles surge el factor "lectura de los Objetivos Fundamentales". Tal y como se mencionó en el primer punto de este mismo documento, si bien los Objetivos Fundamentales contienen habilidades, éstas no pueden ser leídas en forma aislada, ya que es precisamente la relación que tienen con el contenido, contexto, actitudes y otros elementos asociados al aprendizaje, los que marcan verdaderamente los grados de dificultad. Este punto resulta de particular sensibilidad debido a que el currículum vigente ofrece una matemática para la vida, donde los contenidos sin contextos y desconectadas de las habilidades, constituyen aprendizajes carentes de sentido y poco útiles en la vida adulta.

En síntesis, la lectura de Objetivos en busca de la discretización de habilidades, es incorrecta desde la concepción curricular actual, y evidencia diferencias de enfoques entre el currículum aplicado y la evaluación que busca monitorear porcentajes de apropiación.

4. Observaciones relativas a la sección "Estructura de la prueba de selección universitaria de matemática".

En el primer párrafo de las páginas 10 y 11, se deja entender que las preguntas consisten en resolver el problema que se plantea. Pero, como también se dice más adelante, no todas las preguntas entran en la categoría de Resolución de Problemas. Siendo este un término especialmente sensible en Educación Matemática (la Resolución de Problemas), es preferible cuidar su utilización.

El tercer párrafo de la página 11 indica que las preguntas están organizadas por ejes, y dentro de cada eje, por nivel. No es claro si este criterio corresponde a una estrategia que apunta a aumentar la calidad de la evaluación, o es una forma de orden arbitraria y sin pretensiones.

En la descripción de las habilidades cognitivas, no se entiende claramente la separación entre "Comprensión" y "Aplicación". Además, en esta última, se indica "realizar comparaciones a la luz del problema", lo que no es claro a que se refiere. En la

[17] Ver informe "Evaluación de aula en enseñanza básica y media". Componente seguimiento. Unidad de Currículo y evaluación. Ministerio de Educación.

descripción de “Análisis, síntesis y evaluación” se indica “efectuar abstracciones de figuras geométricas, gráficos y diagramas, para resolver problemas”, lo que no es claro su sentido. Además, la resolución de problemas parece quedar instalada en los procesos previos “Comprensión” y en especial “Aplicación”.
Si las definiciones anteriores están descritas utilizando términos y conceptos que generan más de una interpretación, los grados de ambigüedad en la construcción de ítems deberán ser altos también.

5. Observaciones y comentarios adicionales al Marco Teórico de PSU.

Impacto de la prueba en el currículo nacional

La enseñanza de la matemática en el currículum nacional, se enmarca en un enfoque que arranca del contexto en que viven y se desarrollan los estudiantes, así como la íntima relación que existe entre el desarrollo del pensamiento matemático, sus creadores, los problemas y motivaciones culturales, técnicas y científicas que generaron o acompañaron el nacimiento del conocimiento y un aprendizaje significativo. El marco operativo de la PSU, reconoce las limitaciones de los instrumentos elegidos para dar cuenta de estos aprendizajes y procesos. Y, es sabido que “lo que no entra para la prueba” no es importante.

Referencia curricular

Un concepto central en el procedimiento se refiere a la “referencia curricular”. Dice, “la noción de referencia curricular descansa en la necesidad de asegurar una vinculación entre los instrumentos de evaluación y el Marco Curricular vigente…” (Pág. 55). Impresiona la complejidad del procedimiento y esa misma complejidad hace difícil estimar el efecto que el procedimiento descrito tiene en lo que queda y lo que no queda seleccionado para conformar el instrumento definitivo. Sin embargo, la naturaleza cuantitativa de los análisis descritos, hace pensar en la posibilidad de una acción, posiblemente a posteriori, de carácter cualitativo. El razonamiento hecho es el siguiente: se puede considerar que la frecuencia de aparición de una habilidad o conocimiento es indicador de su relevancia, pero podría ser que aprendizajes nombrados una vez o con frecuencia baja, tenga una importancia alta. ¿Se hace uso de opinión experta en este sentido? ¿Es pensable o interesante, incorporar juicio experto en el proceso de selección descrito?

Razonamiento matemático

En la Pág. 9, párrafo 4, nueve, se indica que se trata de una prueba de “razonamiento matemático”, cuyo sentido no queda claro ni se elabora una descripción más adelante. En el mismo párrafo pareciera desprenderse que sería de razonamiento matemático por usar como estímulos para las preguntas, contenidos propios del sector. ¿Significa esto acaso que estímulos relativos a la aplicación de la matemática en la resolución de problemas reales no forma parte de esta prueba? Esto es contrario al espíritu del actual

currículum (y del ajuste) donde se pretende potenciar la matemática como herramienta para comprender el mundo.

Los puntajes de la PSU – MAT

El marco evaluativo para la prueba de Matemática no incluye dentro de sus componentes la tabla de transformación de puntaje. Ésta es dada a conocer a través de otras publicaciones del DEMRE. Aun cuando dicha tabla no forma parte del marco evaluativo, se consideró apropiado plantear algunas preguntas que surgen en relación a ésta: ¿Existe correlación entre los puntajes de la prueba y las notas del colegio?, ¿cómo se explica que un alumno que entregue la prueba en blanco obtenga un puntaje superior al correspondiente a la nota 4,0 del NEM.?

¿Los puntajes de la PSU – MAT identifican a los buenos alumnos?, por ejemplo, un alumno que de las 70 preguntas que contesta cuatro erradas, ¿es inferior en cuánto a conocimiento y competencias matemáticas a un alumno que obtiene las 70 preguntas correctas?; la diferencia de puntajes en la prueba, en ese caso, varía entre 90 y 100 puntos de acuerdo a esta tabla, lo que significa una diferencia en la ponderación final entre 45 y 50 puntos para las carreras que ponderan en un 50% la PSU – MAT; esta situación no se da en las otras pruebas[18].

Ausencias en el documento

Se observan las siguientes ausencias en el documento:

- No se describe en términos englobadores cuál es el constructo medido. Se habla de razonamiento matemático, pero esto no es definido sino a través de una lista de elementos que lo constituyen (ver página 6 - 2° párrafo, página 9 - 5° párrafo y página 98 párrafo final), siendo el principal referente el currículum (OF-CMO).

- No se hace ninguna referencia al tema de la dificultad esperada de las preguntas, ni como se puede graduar esta dificultad. Como existen puntajes mínimos para postular a ciertas Universidades / Carreras, es muy importante saber a que corresponden dichos puntajes mínimos. Estamos aquí en presencia de puntajes de corte y posibles niveles de logros asociados, pero nada de esto se explicita.

- No se hace ninguna referencia a la forma de asignar puntajes las respuestas, o a la prueba.

- No se incluyen ejemplos de preguntas que ilustren lo reseñado en la estructura de la prueba. Sólo se ilustran con ítemes los tipos de preguntas, y los criterios para elaboración de ítems.

[18] Tabla de transformación de puntajes. DEMRE

Capítulo 4. Análisis de marco evaluativo para prueba de Ciencias

1. Introducción

Ante la solicitud de comentar el Marco de Evaluación para la Prueba de Selección Universitaria[19] en nuestro país, parece necesario conceptualizar desde el inicio algunos alcances referidos a lo que entendemos por Marco de Evaluación y cómo este se constituye en el lineamiento general del proceso que comprende.

La necesidad de acompañar o preceder los procesos evaluativos mediante un Marco de Evaluación es relativamente reciente, y en todo caso sigue al interés y desarrollo creciente que ha tenido la evaluación en la esfera educacional, y ha pasado de ser solo una descripción de lo que se evaluará a la formulación de documentos que incorporan una visión integradora de sus distintos componentes.

La declaración de un Marco de Evaluación facilita la comunicación, mediante la puesta en común del significado de los conceptos involucrados, de los distintos procedimientos que se ponen en marcha durante su aplicación, asimismo facilita la comprensión de la información de sus resultados y de propuestas o sugerencias.

Su diseño supone un sistema organizado de los componentes y alcances que dan cuenta de los distintos procesos que se articulan en cada una de las etapas de la evaluación.

Desde esta perspectiva nuestro análisis considera los marcos de Biología, Física y Química como pertenecientes a un solo documento correspondiente a los planteamientos de Ciencias dentro de la PSU.

Es así que, para el análisis que se presenta más adelante, en primer lugar se revisaron diferentes marcos de evaluación de pruebas estandarizadas, y se identificaron aquellos elementos comunes en el desarrollo de sus planteamientos, los que consideraron como referentes para realizar este análisis del Marco de Evaluación de Ciencias de la Prueba de Selección Universitaria.

Adicionalmente, el presente análisis es también llevado a cabo en consideración que la PSU fue concebida como una prueba que vincula el proceso de selección con el logro de los aprendizajes que el currículum define para la Enseñanza Media. En consecuencia, uno de los puntos clave que se analiza a continuación consiste en el

---

[19] Una primera interrogante que se nos presentó al analizar los documentos de la referencia, se relacionó con el titulo de cada uno. Es decir, si los documentos corresponden a Marcos de Evaluación, a Marcos Teóricos o solo corresponden a Análisis del marco Curricular de cada subsector.

Tal como están desarrollados y considerando que se orientan a una prueba de selección, nuestro análisis los asume como Marcos de Evaluación

grado en que el marco de evaluación efectivamente permite evaluar tales aprendizajes. Es decir, este análisis se lleva también a cabo examinando estos marcos en términos del alineamiento curricular que favorecen.

1.1 Aspectos fundamentales de la conceptualización de un Marco de Evaluación

Al revisar diferentes Marcos de Evaluación[20] educacional es posible apreciar en ellos elementos comunes en torno a los cuales desarrollan sus planteamientos. Hemos agrupado estos elementos en cuatro ámbitos que se presentan a continuación, en función de los cuales se analizará posteriormente los marcos de evaluación desarrollados para la PSU. Estos elementos refieren a los aspectos de carácter técnico que una propuesta de este tipo debiera incluir. Sin embargo, la discusión sobre éstos no es ajena al análisis referido al alineamiento curricular. Este último punto es abordado de manera transversal en la discusión de de estos cuatro elementos.

Propósitos y motivos

- Este ámbito permite dar cuenta del para qué y por qué de la evaluación, es decir la finalidad y el motivo que originan la evaluación. Identificar, si es pertinente, las decisiones de política educacional que lo ameritan. Lo que se busca obtener y lo que reflejarán los resultados.
  También es importante presentar las áreas o dominios que se evaluarán; señalar a quiénes va dirigido, es decir quiénes serán los evaluados; mencionar el o los referentes de contrastación considerados al evaluar, por ejemplo el currículum prescrito, las necesidades expresadas por las universidades, u otros.

Justificación teórica

- Este ámbito permite dar cuenta de los lineamientos teóricos en los que se sustenta el Marco de Evaluación, que incluya por ejemplo, una descripción de lo que se entenderá por aprendizaje, alcances sobre las habilidades, etc. Puede ampliar algunas especificaciones sobre los referentes de contrastación usados al evaluar y mencionar los métodos de análisis que se ocuparán. Se requiere fundamentar la matriz de lo que se ha decidido medir.

Metodología a emplear

- Este ámbito permite dar cuenta de cómo se llevará a cabo el proceso. Es necesario señalar cuál será la estructura y componentes de la prueba, lo qué cubrirá cada uno de los instrumentos de evaluación; mencionar los instrumentos que se utilizarán durante la evaluación, su descripción y las dimensiones que evalúa. Tal vez señalar razones de la elección o preferencia de un tipo de instrumento, qué se consigue con ese y no con otro. Conviene presentar cuál será el peso o

[20] Marco Evaluación NAEP 2009; Marco Evaluación PISA 2007; TIMSS 2007 Assessment Frameworks; Marcos teóricos y Especificaciones de Evaluación PIRLS 2006

importancia que tendrá cada área o dominio evaluado dentro del test o instrumento.

Perspectivas a futuro

- Este ámbito se refiere a como los resultados del proceso de evaluación lo retroalimentan, cuánto tiempo se prevé de vigencia de los planeamientos que constituyen el marco, en tanto guía del proceso de evaluación. Asimismo puede dar cuenta o advertir sobre procesos de autodesarrollo que se esperan para algunos componentes del Marco de Evaluación, como consecuencia a su vez de los procesos de retroalimentación. También puede incluir consideraciones de propuestas de actualización, o qué debiera ocurrir para actualizarse.

  Las observaciones relativas a este punto no constituyen el foco principal del presente informe. En consecuencia, los comentarios referidos al mismo no se desarrollan con igual nivel de detenimiento que las referidas a los puntos anteriores.

2. Análisis

Los comentarios y alcances se han organizado de modo que se correspondan con los 4 ámbitos señalados en el punto anterior.

2.1 Propósitos y motivos

Si asumimos que la PSU es una sola entidad evaluativa, parece conveniente que, la Introducción del Marco de Evaluación de la Prueba de Selección Universitaria, junto con presentar los elementos que contendrá, entre ellos las áreas que se evaluarán, abarque los aspectos generales referidos al para qué y porqué de esta evaluación como asimismo señalar las decisiones de política educacional en que se basan.

Si bien, las Introducciones de Biología, Química y Física aparece una referencia a la Prueba de Aptitud Académica vigente hasta el 2003 y se señalan las dimensiones de aptitud y manejo de conocimientos cubiertas respectivamente por la PAA y PCE (prueba de conocimientos específicos) no hay alusión al motivo de este cambio o transición desde la PAA a la PSU. Probablemente se omitió dada la extensa discusión que generó en su momento, no obstante pensamos que la alusión inicial a la PAA y PCE hacen necesaria una explicación del motivo de su cambio a la PSU. Este punto parece ser más consistente con una presentación general.

2.2 Justificación teórica o bases teóricas

Los lineamientos teóricos que se mencionan como bases para el desarrollo de esta evaluación están restringidos a un ordenamiento de los elementos que el currículum define. Si bien esto denota una referencia directa al currículum, se observa poca profundización en las habilidades del pensamiento y proceder científico que debiera evidenciar el estudiante que rinde la PSU, y que son parte fundamental de la propuesta curricular. Conviene destacar la presentación que hace Física en este sentido, la que recoge con mayor amplitud las habilidades que cabe esperar, hayan desarrollado los postulantes.

Al respecto conviene recordar lo que señala el currículum nacional como propósito del área de ciencias:

> Un criterio básico de la selección y organización curricular del sector es que la ciencia es un conocimiento sobre el mundo, que para ser significativo debe ser conectado con la experiencia y contextos vitales de los alumnos. El punto de partida debe ser la curiosidad, ideas propias e intuiciones de los estudiantes; y el punto de llegada, no la mayor cobertura temática posible de una disciplina, sino el entendimiento de algunos conceptos y principios fundamentales de las ciencias, sus modos de proceder, y la capacidad de aplicarlos correctamente[21].

Si bien en el Marco de la PSU se mencionan habilidades, éstas están basadas en la teoría del dominio cognoscitivo planteada por B. Bloom, la que no necesariamente constituye la base teórica del currículum. En consecuencia, el uso de esta teoría no garantiza que las habilidades asociadas al pensamiento y quehacer científicos que este marco promueve sean efectivamente contempladas en la prueba. La interrogante sobre la suficiencia de este referente conceptual de habilidades se pone de manifiesto al considerar tanto los desarrollos conceptuales posteriores a la teoría de Bloom, así como las diferencias que se observan entre ésta y la propuesta que subyace al currículum.

El desarrollo de la psicología cognitiva desde la década del 60 a la fecha, provee de nuevos conocimientos sobre los procesos cognitivos que tienen lugar durante el aprendizaje, por lo que en la actualidad, resulta poco apropiado reducir las habilidades a un verbo. A modo de ejemplo se pueden citar los estudios e investigaciones de Jerome Bruner en la formación de conceptos y de David P. Ausubel en la significatividad de las experiencias de aprendizaje, que han repercutido en la forma como se entienden las estrategias cognitivas, de tal modo que no resulta posible desconocer su aporte.

Uno de los aspectos en los que los marcos de evaluación pueden requerir un mayor desarrollo y profundización refiere al reconocimiento de las características específicas de las habilidades que son propias del trabajo científico. Una visión más elaborada al respecto resulta una condición importante al momento de evaluar el currículum. La conveniencia de esto último se debe a que la conceptualización de habilidades en función de la taxonomía de Bloom no facilita rescatar en la naturaleza que éstas poseen. Esta dificultad se debe a que, sobre la base de esta de esta taxonomía, las habilidades se entienden en función de un verbo específico, es decir, se asume que aquellas

[21] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 125).

habilidades que el currículum contempla están contenidas en una acción planteada en términos genéricos. Al ser abordadas de esta forma, estas habilidades se visualizan como categorías abstractas, independientes del campo de conocimiento y actividad en las que tienen lugar, sin que se distinga el carácter o especificidad que aquellas asociadas a un ámbito determinado puedan poseer.

No obstante, la relación entre las habilidades y su vinculación con la disciplina subyace a las expectativas que plantea el currículum cuando señala que, entre las habilidades que la Educación Media debe fomentar en el plano del desarrollo del pensamiento, se encuentran también "*las de investigación que tienen relación con la capacidad de… revisar planteamientos a la luz de nuevas evidencias; suspender los juicios en ausencia de información suficiente*". Asimismo "*las de análisis, interpretación y síntesis de información y conocimiento, conducentes a que los estudiantes sean capaces de establecer relaciones entre los distintos sectores de aprendizaje; de comparar similitudes y diferencias; de entender el carácter sistémico de procesos y fenómenos*...."[22]

Si bien es cierto que el trabajo realizado para producir el Marco de Evaluación es arduo y extenso, se podría hipotetizar que una falta o ausencia de demandas de las competencias centrales del campo científico produce una reducción de las expectativas de probar las "*capacidades de las personas y del país para utilizar creativamente el conocimiento*" ....y "*las formas de pensamiento típicas de la búsqueda científica*" (las que) "*son crecientemente demandadas en contextos personales, de trabajo y socio-políticos de la vida contemporánea*";... que relevan al propósito del currículo actual de lograr que "*todos los alumnos y alumnas logren en su formación general una educación científica básica*"[23]

Las observaciones arriba planteadas señalan las dudas que derivan del uso de la taxonomía e Bloom como referente para capturar el tipo de aprendizaje que el currículum establece. Sin embargo, existen también otras dudas adicionales que emergen del uso de este referente, y que tienen que ver con la claridad en las distinciones que se pueden obtener por medio de estas habilidades, como por ejemplo:

- ¿Cómo diferenciar si un alumno conoce algo, o bien lo reconoce?
- ¿En qué es diferente identificar de distinguir?
- Si un alumno realiza un experimento ¿cómo pone en evidencia que sabe identificar las variables que intervienen, o si su diseño es adecuado para responder la pregunta de investigación, o si las pruebas de que dispone avalan una determinada conclusión?
- ¿Cuáles son los límites conceptuales para cada verbo-habilidad?

Estas son dudas para las cuales no se pueden desprender respuestas claras desde los marcos de evaluación.

[22] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media (pág. 21).
[23] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media (pág. 124)

2.3 Metodología a emplear

La Metodología utilizada denota un esfuerzo por desarrollar un análisis exhaustivo y riguroso de los contenidos del currículum nacional, de modo que facilite su evaluación.

No obstante, en una mirada general se observa que la inclusión de las numerosas tablas utilizadas para el análisis del currículum dificulta la fluidez de la lectura del documento, y sobre todo, plasmar de manera clara el propósito del mismo. Dada esta situación, podría considerarse la inclusión de las tablas en un anexo, dejando en el Capítulo destinado a la Metodología las citas, las consideraciones y las decisiones que surjan.

En relación con la reestructuración de elementos curriculares y de programas (OF-CMO – Unidades), para generar agrupaciones de contenido en Áreas Temáticas, nos parecen válidas y atendibles las ventajas señaladas en las disciplinas de Química y de Física, toda vez que las agrupaciones están propuestas en función de organizar la evaluación y convergen a la consecución de los OF correspondientes.

Uno de los puntos que merece especial atención en la metodología de análisis curricular consiste en el rol asignado al conteo o cuantificación de los elementos constituyentes del currículum. Si bien la realización de cuantificaciones es necesaria para la elaboración de una prueba, ésta por sí sola no constituye una estrategia que permita evidenciar los aprendizajes que el currículum define. En el caso de la metodología utilizada en los marcos de evaluación, el currículum se analiza en función del conteo de sus elementos, pero sin que este proceso dé cuenta a la vez de un análisis más profundo de la propuesta curricular, y sobre la base del cual se sustente el registro cuantitativo que se lleva a cabo.

Uno de los problemas que trae consigo esta aproximación es que no permite rescatar una visión respecto del carácter específico de los elementos que componen el curriculum. Como consecuencia, se asigna un valor o peso equivalente a algunos aspectos que no son equiparables en términos de los requerimientos o demandas involucradas. Por ejemplo, en el caso de Física, el CMO "Pulsaciones entre dos tonos de frecuencia similar" no puede ser equiparable en términos de extensión y profundidad con el CMO "Distinción entre ondas longitudinales y transversales, ondas estacionarias y ondas viajeras". La asignación de un mismo valor a cada uno de ellos pasa por alto diferencias que requieren ser consideradas para efectos de la evaluación. Una situación similar ocurre en torno al análisis de las habilidades involucradas. Por ejemplo, la habilidad "reconocimiento de concepto" se iguala en peso porcentual a otra habilidad como "análisis y resolución de problemas" lo cual puede producir distorsión dado que éstas involucran diferentes competencias o niveles de logros.

En relación con los ítemes, si bien el Marco de la PSU señala que corresponderán a los conocidos como de Selección Múltiple, no se acompaña el fundamento de esta decisión. Por otra parte, resulta importante la inclusión que se hace de algunas normativas generales para cumplir con las características que deben tener los ítemes a utilizar. Sin embargo no parece igualmente procedente incluir, en el Marco de Evaluación, ciertas formalidades que se avienen más con la operatividad que tiene lugar

durante el proceso de la elaboración de los ítemes, como por ejemplo, el formulario que debe entregar cada elaborador de ítemes.

Con todo, el solo cumplimiento de las normativas de construcción de un ítem no asegura la calidad de este. En otras palabras, si un ítem de selección múltiple consta de estímulo, enunciado y opciones apegadas a las reglas señaladas, no está asegurado "per se" que ese ítem pueda capturar la esencia del aprendizaje propuesto en el currículum.

El contenido, la habilidad y el contexto reflejados por un ítem tienen que estar sustentados en una adecuada interpretación del referente en que se basan, como condición inicial que posibilite la validez de contenido del ítem. En el punto anterior sobre Justificaciones teóricas, se advierte acerca de la debilidad en el desarrollo del tratamiento de las habilidades de pensamiento científico al tiempo que se hace notar el propósito del currículum acerca de la vinculación con el mundo real del los aprendizajes en ciencias, de modo que resultaran significativos para el estudiante.

A continuación se analizan dos de los ítemes que se presentan como ejemplo.

¿Cuál es la geometría de la molécula de CS_2?

A) *Angular.*
B) *Lineal.*
C) *Tetraédrica.*
D) *Trigonal plana.*
E) *Piramidal.*

El ítem está bien formulado, su ortografía es correcta, sus opciones son homogéneas, no hay términos o palabras en el enunciado que sugieran la clave. No obstante ¿mide este ítem un conocimiento relevante a la luz de los aprendizajes que el currículum establece? En otras palabras, ¿es éste un antecedente útil para conocer el grado de dominio de los conceptos, principios y habilidades científicas fundamentales que se busca promover por medio del curriculum?

Junto con ello, cabe también la pregunta por el grado en que este ítem involucra una conjunción contenido-habilidad-contexto tal que puede capturar evidencia de un aprendizaje de calidad. Si bien la ciencia química es parte fundamental del mundo físico que nos constituye, por su naturaleza, está referida a un ámbito en si mismo abstracto, se desenvuelve en el micromundo no visible de átomos y moléculas, por lo que se hace aun más necesaria la presencia de un contexto que facilite la aplicación de los contenidos de modo que no pierdan su significatividad, considerando especialmente la condición de no especialistas de los estudiantes expuestos al currículum.

El ítem del ejemplo, impresiona como sostenido en el recuerdo y memorización de la estructura electrónica de los átomos carbono y azufre participantes de la molécula, y de la formación de los enlaces correspondientes que originarán una estructura de geometría lineal. Por otra parte no hay una contextualización que colabore a situar al estudiante de modo que se sienta convocado a demostrar el uso o aplicación de sus conocimientos.

Otro ejemplo que se presenta, proveniente de otra disciplina, es el siguiente:

¿Cuál de las siguientes hormonas **no** es producida por la placenta?

A) *Estrógenos.*
B) *Progesterona.*
C) *Oxitocina.*
D) *Lactógeno placentario.*
E) *Gonadotropina coriónica*.

¿Es esta memorización lo que queremos como desarrollo en los estudiantes?

Cabe preguntarse, en primer término, ¿dan cuenta estas preguntas del tipo de aprendizajes que se espera desarrollar?, ¿es la memorización de estos contenidos una señal importante para evaluar el dominio de conceptos y habilidades clave? Por otra parte, cabe también preguntarse ¿es posible desarrollar ítemes que manteniendo el contenido de geometría molecular (o de hormonas placentarias), provoque una reflexión, comparación, deducción, inferencia, propuesta de geometría molecular, etc., tal que el estudiante sienta que la pregunta es en realidad un estímulo y un desafío interesante y por otra parte entregue evidencia de que los aprendizajes realizados por ese estudiante lo proveen de herramientas para pensar creativa, crítica y científicamente?

Conviene tener presente que la selección de OF/CMO suponen "*una forma de educación en ciencias que otorga tanta importancia al conocimiento acumulado por las diferentes disciplinas como a sus formas de pensamiento y proceder. Ambos aspectos no deben ser separados. Una enseñanza de la ciencia que se concentra solo en el saber disciplinario acumulado –ciencia como "vocabulario científico"- lleva a un muy bajo entendimiento y ciertamente no al desarrollo de la autonomía intelectual buscada"*[24].

Si se busca la vinculación con los Objetivos Fundamentales[25], el primer ítem señalado podría relacionarse con el OF de 2ª E.M "Relacionar la estructura electrónica del átomo con su capacidad de interacción con otros átomos" y con el CMO: "El enlace químico", específicamente con: "b) Enlaces iónicos, covalentes y de coordinación, c) Descripción de ángulo de enlace, isomería".

El segundo de los ejemplos acá citado, es posible relacionarlo con el OF de 2º E.M: "Apreciar y valorar la interrelación de los aspectos biológicos, afectivos, espirituales, éticos, culturales, sociales y ambientales de la reproducción y desarrollo humano" y con

[24] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 125).
[25] No se señalan las referencias curriculares (OF/CMO) en el Marco de Evaluación presentado

el CMO "Hormonas y sexualidad humana", específicamente con "a) Formación de gamitos, efecto de las hormonas sexuales, ciclo menstrual y fertilización".

Si bien cada uno de estos ítemes, está vinculado con un referente de contenido curricular, cabe preguntarse: ¿siguen éstos la línea u orientación de aprendizaje que el currículum establece?, es decir, ¿miden éstos un conocimiento relevante?, ¿se evalúa por medio de ellos el desarrollo de conocimientos y habilidades fundamentales y generativas? Y junto con ello, ¿miden lo que el estudiante sabe hacer con este conocimiento?, ¿está conectado con la experiencia y contextos vitales de los alumnos, para asegurarse de que sean significativos?[26]

En las Consideraciones generales para el desarrollo de las distintas etapas en la construcción de ítemes de calidad, se postula la necesidad de tener un "acabado conocimiento de los contenidos curriculares vigentes y de las habilidades cognitivas que serán evaluados"[27].

Sin duda, es este un requisito que invita a una reflexión y análisis más profundo que se oriente a esclarecer cómo y de qué manera es posible incrementar el desarrollo de ítemes que consideren no solo conocimientos aislados sino las habilidades cognitivas asociadas que conlleva el proceso de enseñanza de las ciencias, cuyo estudio por la psicología del aprendizaje ha experimentado un progresivo avance y constituye un claro planteamiento de la formulación curricular cuando plantea:

Un supuesto central del sector es que el aprendizaje de la ciencia debe ser un proceso activo en el cual la investigación y la resolución de problemas ocupan lugares centrales; se sostiene, además, que estas actividades de investigación y experimentación son decisivamente más ricas en términos de aprendizaje, si se las desarrolla en contextos donde se conjuguen elementos de historia de la ciencia, perspectivas sociales y personales sobre sus usos, y aplicaciones tecnológicas contemporáneas".[28]

Al considerar la relevancia que éstos aspectos debieran tener en la prueba, emerge la interrogante relativa a la posibilidad de de utilizar otras alternativos o tipos de preguntas por medio de las cuáles éstos puedan ser evaluados. ¿Cuál es el criterio que indica que no es posible el desarrollo de ítems abiertos?, ¿es posible de desarrollarlos en test de esta magnitud para levantar la medición efectiva de aprendizajes relevantes? Si no es así, faltaría fundamentar mayormente el por qué de esta situación y más aún no solo desde la perspectiva de ítems abiertos sino que de otras alternativas que eventualmente pudieran levantar evidencias más claras de los alcances y logros de aprendizajes relevantes por parte de los estudiantes

Finalmente, cabe esperar que en la metodología se de cuenta de otros procesos relacionados con la aplicación de los ítemes como los siguientes: fases de experimentación de los ítemes; validación del constructo que sirve de base para la selección de lo evaluado; características de la aplicación, tales como: numero de formas, selección de los ítemes de comparabilidad (equating), condiciones de la

[26] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 125).
[27] DEMRE. Marco Teórico de la PSU del Sector Ciencias Naturales. Subsector Biología. (Pág. 99).
[28] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (Pág. 125).

aplicación, y cómo se harán el procesamiento de los datos y el análisis de los resultados.

2.4 Perspectivas a futuro

Como se señaló anteriormente, este punto en particular no constituye el foco principal del presente análisis. Sin embargo, hay algunos comentarios de orden general que se considera pertinente incluir.

En las declaraciones iniciales está subyacente la idea que los resultados servirán de base para la postulación a la Universidad. No se señala, ni explícita ni implícitamente, si existe alguna vía o canal que permita a las Universidades retroalimentar el proceso con sus expectativas, sugerencias o peticiones, acerca de las competencias que requieren haber desarrollado quienes postulan a sus aulas. Tampoco hay una estimación de un período de tiempo durante el que este marco esté vigente, como tampoco se señala un tiempo o circunstancias cuya ocurrencia pudieran ameritar su revisión y ajuste. Todos estos son elementos que se esperaría encontrar en un marco de evaluación, con el objeto de contar con una visión del desarrollo y perfeccionamiento de la prueba a futuro.

Capítulo 5. Análisis de marco evaluativo para prueba de Historia y Ciencias Sociales

Introducción

Este informe contempla dos partes. En la primera se comenta el "Marco Teórico Curricular" de la PSU de Historia y Ciencias Sociales, con especial énfasis en el análisis del currículum nacional que el DEMRE realiza en orden a construir el instrumento de medición, y en la matriz curricular que se utiliza para organizar la prueba. La segunda parte consiste en un anexo con los errores detectados en la formulación de preguntas y en la justificación de las correcciones en "Resolución Facsímil. Prueba Historia y Ciencias Sociales" de 2007, liberada por el DEMRE en 2008 y publicada en su sitio web.

Comentarios al "Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales"

El documento "Marco Teórico Curricular" consiste en un análisis del currículum nacional de enseñanza media en el sector de Historia y Ciencias Sociales a partir del cual se elabora la PSU de Historia y Ciencias Sociales como instrumento de medición.[29] Este documento se desglosa en nueve capítulos.

El primero consiste en una revisión de los antecedentes de la PSU, a saber: la prueba de conocimientos específicos de Historia y Geografía y la prueba de conocimientos específicos de Ciencias Sociales, que estuvieron vigentes entre 1985 y el 2003, y la Prueba de Selección Universitaria (PSU) de Historia y Ciencias Sociales según la matriz utilizada hasta el proceso de admisión 2007. En este primer capítulo se da cuenta detallada de la estructura de cada una de estas pruebas y de los principales cambios entre las antiguas pruebas de conocimientos específicos y la actual PSU.

En el segundo capítulo, se analiza específicamente el currículum del sector de Historia y Ciencias Sociales para la enseñanza media, desglosándolo en categorías que se presentan como conducentes a su transferencia hacia un instrumento de medición.

En el tercer capítulo, se expone la noción de "referencia curricular" tal y como se consensuó en la Mesa Escolar del 2002 (entre el DEMRE, el Mineduc y el Comité

[29] DEMRE (2006). Marco teórico curricular. Prueba de selección universitaria Historia y Ciencias Sociales (pág. 6).

Asesor del Consejo de Rectores), se precisan las categorías según las cuales opera el concepto de referencia curricular, y cómo ellas se articulan para la construcción de las preguntas.

En el cuarto capítulo se expone la Matriz Curricular utilizada para la construcción de la prueba entre los procesos de admisión 2004 hasta el del 2007, mientras que en el capítulo quinto se da cuenta de la nueva Matriz Curricular utilizada a partir del proceso de admisión 2008. El capítulo seis evidencia los criterios utilizados para la elaboración de los ítems que conforman la prueba, mientras que en el séptimo capítulo se exponen los supuestos teóricos sobre los que se fundamenta esta medición. Finalmente, los dos últimos capítulos corresponden a las conclusiones del documento y a la bibliografía.

Tras analizar con detención este "Marco Teórico Curricular" elaborado por el DEMRE, hemos llegado a la preocupante conclusión que l**a PSU de Historia y Ciencias Sociales se estructura sobre una lectura errada del currículum del sector, lo que se traduce en un instrumento de medición que en los hechos no evalúa al referente curricular nacional.** Esto no es menor considerando el impacto de esta medición sobre el sistema escolar. De esa forma, la PSU de Historia y Ciencias Sociales no sólo limita las posibilidades y los sentidos del currículum, sino que **termina constituyéndose en un currículum paralelo.**

Lo anterior se debe, en lo medular, a la forma cómo, a la hora de elaborar el instrumento de medición, son entendidas las categorías que articulan al currículum, particularmente los Objetivos Fundamentales, así como también la manera en que es comprendida la lógica según la cual se construye la matriz curricular vigente, sobre la que se estructura la prueba.

Para argumentar los juicios anteriores centraremos nuestros comentarios en el "Análisis del Marco Curricular del sector de Historia y Ciencias Sociales" presentado por el DEMRE y en la "Matriz curricular para la elaboración de ítemes en la PSU Historia y Ciencias Sociales", correspondientes a los capítulos 2 y 5 del documento que estamos comentando.

a. Sobre el "Análisis del Marco Curricular del sector de Historia y Ciencias Sociales"

El propósito de este capítulo es analizar el currículum del sector, "de acuerdo a un conjunto de categorías que acrediten su transferencia a un instrumento de medición de los Contenidos Mínimos Obligatorios (CMO) y que integre las habilidades representadas por los Objetivos Fundamentales (OF) para el sector de aprendizaje".[30]

Llama la atención el tipo de análisis y lectura que se hace de los Objetivos Fundamentales y de los Contenidos Mínimos Obligatorios, dado que la relación entre Objetivos Fundamentales y Contenidos Mínimos se aborda desde una perspectiva

[30] DEMRE (2006). Marco teórico curricular. Prueba de selección universitaria Historia y Ciencias Sociales. (pág. 23).

eminentemente cuantitativa, contabilizando la cantidad de OF y CMO por nivel escolar (lo que se grafica en tablas y gráficos de barras) y verificando a cuántos CMO se refiere cada OF. Este tipo de aproximación a la relación de los OF con los CMO se aleja de la lógica con la que se construyó el curriculum nacional. En este sentido es fundamental precisar que los Objetivos Fundamentales son definidos en el currículum nacional como "las competencias o capacidades que los alumnos y alumnas deben lograr al finalizar los distintos niveles de la Educación Media y que constituyen el fin que orienta al conjunto del proceso de enseñanza-aprendizaje."[31] En esta lógica, los OF (que consideran conocimientos, habilidades y disposiciones) son la categoría curricular de mayor relevancia, ya que otorgan sentido al tratamiento de los CMO.

Sin embargo, el documento del DEMRE hace una lectura radicalmente distinta de esta categoría curricular, al entenderla sólo como "indicadores de las habilidades que se requieren".[32] Como consecuencia de ello, se analizan los OF tan sólo marcando el verbo que los articula, para posteriormente agrupar ese listado, sólo de verbos, entendidos como "acciones pedagógicas", siguiendo para ello la taxonomía de Bloom que da origen a cuatro categorías de habilidades, es decir: reconocimiento; comprensión y aplicación; análisis, síntesis y evaluación; y por último habilidades valóricas que no son medidas por la PSU. Al respecto, también llama la atención que los OF sean comprendidos como acciones pedagógicas del docente siendo que lo propio de un objetivo fundamental es describir las competencias de los alumnos.

Ejemplificando lo anterior, en segundo año medio el currículum prescribe como uno de los objetivos fundamentales "Comprender que el conocimiento histórico se construye a base de información de fuentes primarias y su interpretación y que las interpretaciones historiográficas difieren entre sí, reconociendo y contrastando diferentes puntos de vista en torno un mismo problema.", siguiendo el documento del DEMRE de este objetivo fundamental solo se rescata el verbo, es decir "comprender" el que es conceptualizado como acción pedagógica. En cuarto año medio el currículum define como uno de los objetivos fundamentales "Entender la complejidad de algunos de los grandes problemas sociales del mundo contemporáneo, como son la pobreza y el deterioro medio ambiental; comprender que su resolución no es simple y que implica la acción conjunta de diversos actores sociales; valorar la solidaridad social y a importancia del cuidado del medio ambiente." Según la metodología utilizada por el DEMRE este objetivo se reduce a "Entender, comprender y valorar". Es evidente que la metodología utilizada se traduce en que se pierde la función curricular de los OF y que los CMO carezcan de orientación en su tratamiento. A nuestro juicio este es uno de los principales factores que permite explicar la distancia observada entre el currículum nacional y los ítemes del facsímil liberado.[33]

Así, aspectos medulares de la propuesta curricular como el énfasis en la historia reciente y la sociedad contemporánea tienen escasa presencia en la prueba analizada comparándolo con el espacio que se le da a los períodos más pretéritos de la historia.

[31] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (pág. 7).
[32] DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 27).
[33] Mineduc (2005). Objetivos Fundamentales y Contenidos Mínimos Obligatorios de la Educación Media. (pág. 104 y 106). DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 75 y 77).

Otro ejemplo del mismo problema es que una dimensión central de la propuesta curricular, como el contraste entre interpretaciones historiográficas esté absolutamente ausente de la medición.

Por tanto, el problema es que el DEMRE asume que da cuenta de los OF del currículum al cruzar los Contenidos Mínimos con las habilidades definidas siguiendo la taxonomía de Bloom, sin considerar que los OF no son ni habilidades ni acciones pedagógicas, sino que aprendizajes que los alumnos y alumnas deben lograr, los cuales se expresan en conocimientos, habilidades y disposiciones. Esta inconsistencia tiene un doble efecto. Por una parte, se desvirtúan los OF y se transforman en lo que no son, y por otra, se trata a los CMO sin un criterio que defina su orientación y extensión, siendo que en el Currículum Nacional, su sentido y tratamiento está dado desde los OF. En consecuencia, ello deriva en la medición de detalles que a la luz del currículum son intrascendentes,

Junto con ello, llama poderosamente la atención que la taxonomía de habilidades utilizada sea la misma que se usó anteriormente para las pruebas de conocimientos específicos de Historia y Geografía y de Ciencias Sociales vigentes entre 1983 y el 2003, cuando el referente curricular nacional no consideraba OF. El problema se agrava dado que la taxonomía de Bloom no tiene relación con la arquitectura del currículum vigente.

En este contexto, el extenso análisis cuantitativo de los OF y CMO, y las múltiples formas en que éste se representa, pierde sentido, ya que en estricto rigor los Objetivos Fundamentales, en su espíritu y propósito original no son considerados en la medición, lo que inevitablemente se traduce en un profundo distanciamiento entre el currículum nacional y la PSU.

b. Sobre la "Matriz curricular para la elaboración de ítemes en la PSU Historia y Ciencias Sociales".

Siguiendo con la revisión del documento, hasta el proceso de admisión 2007 la matriz curricular se estructuró en torno a cuatro ejes temáticos, correspondientes a los objetivos y contenidos de los cuatro años de la enseñanza media:
Región y País
Raíces históricas
Universalización de la cultura
El mundo de hoy

A partir del proceso de admisión 2008 se producen importantes cambios en la estructura del modelo de referencia curricular, los que a juicio del DEMRE se justifican porque "la experiencia en la elaboración de los instrumentos 2004, 2005, 2006 y 2007, muestra la estrecha relación existente entre algunos contenidos y/o unidades de

distintos ejes temáticos.". Debido a lo anterior "… la matriz de referencia curricular que se propone establece 3 Ejes temáticos." [34]

1. Espacio geográfico nacional, continental y mundial
2. Raíces históricas de Chile
3. El legado histórico de Occidente.

Junto con ello en el mismo documento se señala que este cambio "marca gran diferencia con la estructura actual lo que podría provocar un gran impacto en la opinión pública, profesores y alumnos."[35] Sin duda la forma en la que se construye la PSU impacta fuertemente sobre el sistema escolar condicionando el curriculum implementado.

b.1 La incongruencia entre la matriz utilizada y el currículum en Historia y Ciencias Sociales.

El principal problema, y el de implicancias didácticas más graves, es la incongruencia entre la definición de los ejes y los sentidos del currículum. En la introducción del currículum de Historia y Ciencias Sociales para la enseñanza media se señala que éste tiene como propósito "desarrollar en los estudiantes conocimientos, habilidades y disposiciones que les permitan estructurar una comprensión del entorno social y les oriente a actuar critica y responsablemente en la sociedad." Junto con ello se declara, que "la propuesta curricular parte del supuesto que cada uno de las disciplinas que conforman este sector tienen un aporte específico que entregar, el cual constituye un fundamento básico en la formación de los estudiantes. No obstante se postula como necesario lograr una enseñanza integrada de estas disciplinas para asegurar que los estudiantes no se queden con una visión fragmentada de la realidad social".

Considerando el impacto de la PSU sobre el sistema educativo y sobre el currículum implementado nos parece que el modelo de referencia curricular definido limita las posibilidades del currículum en tanto sus ejes apelan sólo a las dimensiones espacial y temporal de todos los problemas o procesos, lo que se verifica al profundizar sobre la justificación disciplinaria de cada uno de los ejes, quedando fuera o perdiendo visibilidad los objetivos y contenidos relativos al funcionamiento político y económico de la sociedad contemporánea. Así **un currículum de historia y ciencias sociales, cuyo principal propósito es entregar a los estudiantes una comprensión del entorno social que les permita actuar en él, queda reducido a un currículum de historia y geografía.**

[34] DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 129).
[35] DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 145).

b.2 Las inconsistencias de los ejes al interior de la matriz y sus implicancias sobre el currículum implementado.

A nuestro juicio, este nuevo modelo de referencia curricular utilizado a partir del proceso de admisión 2008, presenta problemas estructurales, más allá de su incongruencia con el currículum. En lo medular, aparecen profundas inconsistencias entre las definiciones de cada eje temático, los CMO que se agrupan bajo cada eje, y lo que efectivamente se evalúa en ellos.

A modo de ejemplo, al presentar el eje de espacio geográfico nacional, continental y mundial se señala que en él se "analiza el aspecto geográfico desde el entorno inmediato, el del país, el de América Latina y del mundo contemporáneo", esto en "un eje en que prima un mismo enfoque disciplinario: la dimensión espacial de los principales componentes geográficos existentes en el entorno local, nacional, continental y mundial, siguiendo una secuencia espacial de lo particular a lo general."[36] De lo anterior se desprende que se trata de un eje orientado disciplinariamente desde la geografía y centrado en el análisis espacial de los fenómenos estudiados. Sin embargo, el mismo documento señala que bajo este eje se agrupan "elementos geográficos, económicos, del uso de las tecnologías, de los proceso de globalización, y de la participación de Chile en América latina y el mundo".[37] En este sentido se hace evidente que tanto la concepción como la nominación del eje no permiten dar cuenta de todos los elementos que se pretende evaluar en él. Este problema se evidencia al verificar que contenidos referidos al problema económico (1.4.3), al problema de coordinación económica (1.4.4.), a la revolución tecnológica e informática (4.1.4), se evalúan desde un eje definido desde la geografía y justificados desde el estudio de la dimensión espacial de los principales componentes geográficos. Aún más complejo es entender cómo contenidos referidos a tratados internacionales sobre derechos humanos e igualdad de oportunidades para mujeres y hombres (4.4.3) y referido al análisis de la sociedad contemporánea (4.1.5), estén igualmente en este eje. [38] Una lectura posible es que lo que la PSU releva es la dimensión espacial de estos problemas y fenómenos, lo que a nuestro juicio es una errada lectura del Currículum Nacional. Si los contenidos recién mencionado tienen un espacio privilegiado en la propuesta curricular nacional es porque son considerados relevantes tal y como están prescritos en el currículum y no solo en su expresión espacial.

En el eje Raíces Históricas de Chile, se asume que las raíces históricas del país se limitan al desarrollo histórico nacional y no se relacionan con el desarrollo histórico de Occidente, ya que éste se encuentra bajo otro rótulo. Esto transforma a la historia de Occidente en una comprensión ajena a las raíces históricas nacionales, lo que se traduce en una toma distancia de los sentidos del curriculum nacional e impacta en el tipo de preguntas que tratan los procesos históricos occidentales. Además, en este eje se

[36] DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 129).
[37] DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 129).
[38] Los códigos entre paréntesis corresponde a los definidos por el DEMRE para identificar los contenidos.

incorporan contenidos referidos a la organización política vigente, los que difícilmente pueden agruparse bajo la noción de raíces históricas, ya que su énfasis, según el currículum, debe estar en las definiciones conceptuales y en su comprensión aplicada a la vida ciudadana. Por lo mismo, en forma similar al caso anterior, los principales problemas aparecen al contrastar las definiciones disciplinares que justifican al eje y los contenidos que éste efectivamente aborda.

En la justificación del eje se lo define como "el mismo enfoque disciplinario: la dimensión temporal de las distintas etapas y acontecimientos de la historia nacional hasta llegar a la institucionalidad cívica de Chile en la actualidad, que es consecuencia directa del desarrollo histórico del país inserto en el mundo occidental…"[39] Se desprende de lo anterior que el enfoque disciplinario al que se alude es la historia, y que bajo esta disciplina, y en tanto producto histórico, se estudia la institucionalidad vigente. Pero, al revisar los contenidos que se han agrupado bajo este eje, lo que efectivamente se evalúa son las características de la institucionalidad vigente y no su proceso de construcción histórica. El estudio de la institucionalidad vigente en el marco del Currículum Nacional supone una aproximación conceptual para profundizar en sus características y funcionamiento a fin de problematizar las implicancias que tiene para la vida de los ciudadanos.

Nuevamente consideramos que tanto el enfoque como el nombre del eje no permiten dar cuenta de lo que efectivamente se agrupa en él.

Sumado a lo anterior aparecen problemas de otro orden, en tanto en este eje se "analiza el estudio de la historia de Chile desde las civilizaciones precolombinas americanas y su proyección hasta las últimas décadas del siglo XX".[40] Así, se está entendiendo a las civilizaciones precolombinas como parte constituyente de la historia de Chile, lo que es reduccionista y desconoce que su existencia es anterior al proceso histórico mediante el cual se constituye la nación; por cierto, en el caso de las civilizaciones mesoamericanas es un error más evidente aún subsumirlas como parte de la historia de Chile.

En síntesis, el modelo de referencia curricular efectivamente considera la totalidad de los CMO (no así de los OF) de la enseñanza media en los tres ejes que lo constituyen. Sin embargo, como se argumentó con anterioridad, existen importantes inconsistencias entre las definiciones conceptuales y disciplinarias de los ejes y los contenidos que los conforman.

Finalmente, creemos que un Modelo de Referencia Curricular que esté en sintonía con el currículum nacional debería considerar ejes que apelaran tanto a la perspectiva histórica de la sociedad, al análisis del espacio geográfico como a las dinámicas de la organización política democrática y el desarrollo entendiendo que no son compartimientos estancos sino que se encuentran estrechamente relacionados entre si. Junto con ello dicho modelo de referencia curricular debería estar orientado a relevar los vínculos de la historia con el presente y la realidad vivida, centrarse en los conceptos más que datos puntuales, otorgar un lugar privilegiado a las habilidades de

[39] DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 133).
[40] DEMRE (2006) Marco Teórico Curricular. Prueba de Selección Universitaria Historia y Ciencias Sociales. (pág. 133).

análisis e interpretación, así como a la confrontación de interpretaciones historiográficas y sociales. Un modelo de estas características permitiría agrupar los OF/CMO con mayor coherencia permitiendo que su evaluación esté en sintonía con el tratamiento que de estos tópicos propone el currículum vigente.

ANEXO Informe de Marco Evaluativo en Historia y Ciencias Sociales

Errores detectados en la formulación de preguntas y en la justificación de las correcciones en "Resolución Facsímil. Prueba Historia y Ciencias Sociales" de 2007, liberada por el DEMRE en 2008 y publicada en su sitio web.

En el análisis de este documento se han detectado múltiples errores tanto en la formulación de las preguntas, como en la formulación de las alternativas, en la decisión sobre la respuesta correcta, y en la justificación de la corrección que ha entregado el DEMRE.

De las 75 preguntas de la prueba, hay, a nuestro juicio, 14 con errores, sea en la formulación de las preguntas o en las alternativas planteadas, correspondiente al 18.6% de la prueba. Mientras 8 preguntas presentan problemas en los comentarios a las preguntas realizados por el DEMRE.

A continuación se entrega el detalle de los errores detectados organizados según los ejes en los que se estructura la prueba.

a. "El espacio geográfico nacional, continental y mundial", 24 preguntas.

1.- En este eje, el que en el marco de referencia curricular es justificado disciplinariamente desde la geografía se formulan cinco preguntas que corresponde a la disciplina de la economía (preguntas 18 a 22), una pregunta correspondiente a historia contemporánea (pregunta 23) y una pregunta correspondiente a principios de la democracia (pregunta 24). Esto ratifica nuestro juicio respecto a los problemas de enfoque y consistencia en la concepción de este eje. Suponemos que esto debe ser llamativo no solo para el equipo ministerial sino también para docentes y estudiantes.

2.- A nuestro juicio existen errores en la formulación de las preguntas, en las alternativas y/o en la justificación de las respuestas correctas en los ítemes que se detallan a continuación:

2.1.- Hay cuatro preguntas (números 8, 12, 20 y 24) que contienen errores tanto en el enunciado como en las alternativas.

Pregunta 8: La alternativa III "renovación urbana" es un concepto de urbanismo que no puede ser comprendido únicamente como una medida para enfrentar el proceso de crecimiento y extensión en algunas ciudades chilenas. En ciertas ciudades como Valparaíso por ejemplo, la renovación urbana no busca enfrentar el crecimiento de la ciudad sino su recuperación patrimonial. Además este concepto no tiene referente curricular.

Pregunta 12: Hay dos alternativas correctas, la c y la e. La alternativa c también es una respuesta correcta porque los países pobres han realizado múltiples esfuerzos por detener e impedir la emigración hacia países más ricos. Medidas como la reconversión económica y las políticas de subsidio, son iniciativas propiciadas por algunas naciones para frenar la migración.

Pregunta 20: Mientras que las alternativas dan cuenta de un sistema económico centralizado, el enunciado podría estar dando cuenta de un sistema mixto al decir "autoridad generalmente pública", "trasmitir directivas económicas", y "realizar objetivos económicos".

Pregunta 24: Tiene un error al considerar correcta la alternativa I según la cual la igualdad de oportunidades laborales para mujeres y hombres en Chile sería necesario porque más de la mitad de la población nacional son mujeres. Por cierto, da lo mismo cuál sea la proporción de la población femenina en el total nacional para hacer valer este principio de igualdad.

2.2.- Errores en la fundamentación de las preguntas y alternativas:

Pregunta 6: está mal fundamentado el error de la alternativa e) ya que ésta es incorrecta porque no se puede inferir de los datos de la tabla y no por las características que tenga la tasa de fecundidad actualmente en Chile.

b. "Raíces históricas de Chile", 27 preguntas.

1.- En este eje que según el DEMRE apela a las raíces históricas del país llama la atención la presencia de cinco preguntas de teoría democrática y sistema político chileno (preguntas 47 a 51)

2.- Errores en la formulación de las preguntas, en las alternativas y/o en la justificación de la respuesta correcta:

2.1.- Hay siete preguntas que contienen errores tanto en el enunciado como en las alternativas (números 28, 32, 35, 38, 39, 40, 46):

Pregunta 28: Para que la pregunta efectivamente mida análisis, síntesis y evaluación, la respuesta correcta debe desprenderse del texto citado. Sin embargo, la alternativa III considerada correcta no se desprende del texto, por lo tanto debería ser considerada errónea. De lo contrario la pregunta estaría midiendo conocimientos adquiridos y el texto citado sería irrelevante. Adicionalmente, está incorrectamente citada la referencia del texto, pues no menciona al autor del capítulo correspondiente sino al editor del volumen.

Pregunta 32: La alternativa presuntamente correcta (e) está errada porque el panamericanismo es un proyecto propio del siglo XX e incluye a los Estados Unidos. De modo que es imposible que Bolívar considerara "apoyar férreamente la filosofía del panamericanismo". Por cierto, además, el panamericanismo no es una filosofía. Más cerca de lo correcto, aunque no enteramente, sería la influencia del sistema federal norteamericano.

Pregunta 35: Hay errores en el enunciado y en las alternativas. En el enunciado, es incorrecto afirmar que en Chile hubiesen habido "luchas político-religiosas" pues los conflictos políticos que hubo en estas materias nunca llegaron a convertirse en "luchas" como las guerras civiles en Colombia, por ejemplo. También es incorrecta la alternativa II según la cual estas "luchas" se habrían originado porque la

Iglesia quería evitar la separación Iglesia-Estado. Lo cierto es que en los conflictos político-religiosos del siglo XIX no estuvo entre los propósitos de los liberales de gobierno ir a la separación de la Iglesia y el Estado, por tanto, mal podría la Iglesia haber luchado para evitarlo. Por tanto, la alternativa b) no es correcta como lo estima el DEMRE sino que lo es la alternativa a).

Pregunta 38: Está mal construida la alternativa b, que aparece como errónea, en la cual se afirma que la precaria situación de los obreros en las salitreras se debió a que la doctrina social de la Iglesia no penetró entre los patrones. Esta alternativa está mal construida porque esta afirmación es un contrafactual que no se puede probar ni refutar. Al incluir esta alternativa como opción incorrecta, la pregunta adquiere sesgo ideológico.

Pregunta 39: Está errada, pues considera incorrecto que las fichas salitreras se desvalorizaban al entrar al mercado financiero (alternativa a) siendo que las fichas de una salitrera se podían canjear depreciadamente en el comercio. Al circular como medio de pago depreciado se podría considerar que entraban al mercado financiero. Por lo tanto, lo que se considera correcto, que la ficha sólo tenía valor en la oficina que la emitía, la alternativa b), está incorrecta en la medida en que se la confronta con la alternativa a).

Pregunta 40: Está errada la información del enunciado: Federico Errázuriz E. gobernó 5 años y no 4, pues cuando muere ya está por terminar su período y ya había sido elegido su sucesor. Además del cuadro con el número de gabinetes y ministros por Presidente no se puede deducir la alternativa d, que es la que aparece como correcta. La pregunta aparece como evaluando análisis y en ese caso la información tendría que desprenderse del cuadro. Sin embargo, sólo se puede contestar esta pregunta a partir de un conocimiento muy convencional de historia de Chile, que se contradice con muchos planteamientos actuales sobre las características del parlamentarismo en Chile; por ejemplo Heise (1974) sostiene que el régimen parlamentario comienza en 1861 y prueba que durante el período posterior a 1891 se mantuvieron políticas gubernamentales de largo plazo, lo que contradice la justificación de la pregunta. Esa misma visión añeja del parlamentarismo se repite en la fundamentación de la pregunta 44.

Pregunta 46: La pregunta está mal construida pues la alternativa I es imprecisa. ¿Por qué dos millones de toneladas de producción de hierro sería "mínima"? Para considerar correcta la alternativa I habría que pensar que la Gran Depresión fue en 1920 y no en 1930.

2.2.- Hay cuatro preguntas que tienen errores en la fundamentación de la respuesta y de las alternativas (números 31, 36, 40, 42):

Pregunta 31: Está mal fundamentado el error de la alternativa d) porque una elite liberal habría sido de todas maneras contraria a los realistas, pero en el proceso de emancipación la elite criolla no es aún liberal, sólo una pequeña fracción de ella, la más doctrinaria, es la que va siendo cada vez más liberal.

Pregunta 36: Está errada la fundamentación de la opción I puesto que los Conservadores no son el partido de gobierno en los decenios, sino que este partido se crea como tal a partir de la división de los partidarios del gobierno entre nacionales y clericales. Es incorrecto decir que se produjeron "violentas divisiones" entre liberales dando origen al partido Radical. El partido Liberal Democrático representa a los balmacedistas después de 1891, no sólo a los seguidores de Vicuña Mackenna entre 1875 y 1886. En la fundamentación de la opción II es errado sostener que los liberales agrupaban a "industriales y mineros" pues había muchísimos terratenientes liberales,

así como había "mineros" conservadores. De hecho, contrariamente a lo que afirma el DEMRE, la Sociedad de Fomento Fabril nació de una iniciativa del gobierno, el que solicitó a la Sociedad Nacional de Agricultura que impulsara la creación de esta nueva asociación.

Pregunta 40: Es erróneo afirmar que entre 1891 y 1925 hubo en Chile inestabilidad política. La sucesión presidencial en 1910 prueba justamente lo contrario: la enorme estabilidad política existente.

Pregunta 42: Está errada la fundamentación del error de la alternativa e). El voto censitario se termina de hecho en Chile con la ley electoral de 1874 que establece la presunción de derecho de que quien sabía leer y escribir tenía la renta necesaria para votar, y no tiene nada que ver con el ascenso de los sectores medios como se afirma por el DEMRE.

c. "El legado histórico de Occidente", 24 preguntas.

1.- En este eje sobre el legado histórico de Occidente, se ha enfatizado la historia premoderna, con las preguntas 52 a 61, dejando sólo seis preguntas para la historia contemporánea (70 a 76), lo que distorsiona los sentidos del curriculum.

2.- Errores en la formulación de las preguntas, en las alternativas y/o en la justificación de la respuesta correcta:

2.1.- Hay tres preguntas que contienen errores tanto en el enunciado como en las alternativas (números 53, 58, y 75):

Pregunta 53: Error en la alternativa d) porque dos civilizaciones pueden ser contemporáneas y a la vez ser desconocidas entre sí.

Pregunta 58: Esta pregunta no tiene referente curricular además contiene errores puesto que en el sistema jurídico chileno la pena de muerte está contemplada en la Constitución y en el código de justicia militar.

Pregunta 75: Está mal formulada la opción I porque no se puede deducir del texto presentado en el enunciado.

2.2.- Hay tres preguntas que tienen errores en la fundamentación de la respuesta y de las alternativas (números 56, 58, 59):

Pregunta 56: Es incorrecta la justificación del error de la alternativa d) pues en el sistema político chileno la elección directa de los gobernantes no implica una elección indirecta de los funcionarios de gobierno; en el sistema chileno actual no existe la elección indirecta.

Pregunta 58: Está errada la fundamentación de la opción I porque de la publicidad de la ley no se puede deducir que ésta protege a las personas.

Pregunta 59: Dice "Baja Edad Media", debe decir "Alta Edad Media". Además se define inadecuadamente a la sociedad estamental.

Appendix Z. *Ajuste Curricular – PSU*

The Curriculum Unit met with DEMRE in 2010 to establish working groups to address the propositions found in the aforementioned 2009 report. The challenges of the PSU addressing the ongoing curricular changes were addressed by the Curriculum Unit in these meetings. A summary of the key themes of these meetings is included in a PowerPoint presentation, *Ajuste Curricular – PSU,* incorporated in the following pages of this appendix.

The appended document is a summary of a PowerPoint presentation made by the Curriculum Unit in its 2010 meeting with DEMRE concerning the PSU and the National Curriculum.

GOBIERNO DE CHILE

Ajuste Curricular

PSU

Ministerio de Educación, Septiembre 2010

GOBIERNO DE CHILE

Objetivos

- Tomar decisiones respecto a camino a seguir para lograr alineamiento curriculum PSU
- Informar sobre los cambios en las asignaturas provocados por el ajuste.
- Conversar sobre el cronograma de Curriculum 2010-2014.

Contexto

- 2010: entrada en vigencia de ajuste curricular desde 5° básico a 1° año medio, y gradualmente entrarán en vigencia los demás cursos.
- Se reelaboraron programas para estos cursos . 2010 (están en el CNE)
- Se entregaron textos con ajustes desde 1º básico a IIº medio.

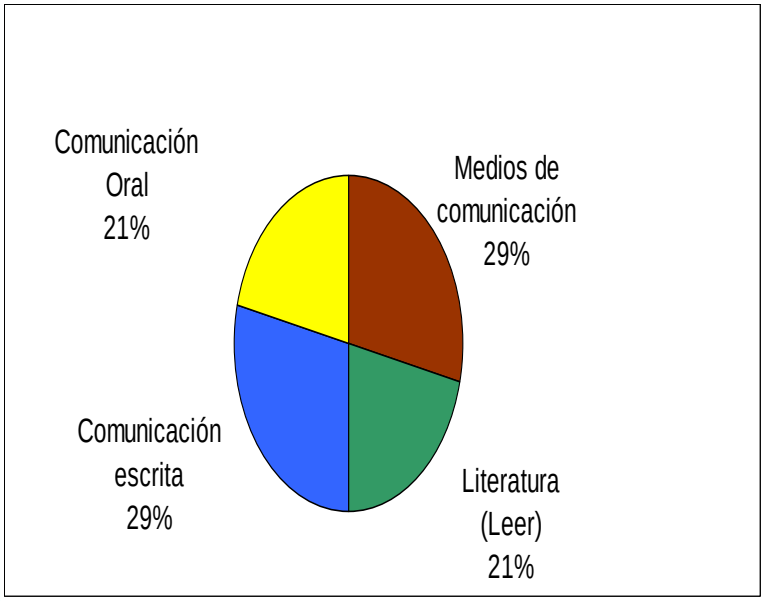
Lenguaje
Principales cambios

	2005	2009
Ejes	•Comunicación Oral •Comunicación Escrita •Literatura •Medios Masivos de la comunicación	Se articula la Básica y la media con los mismos ejes: •Comunicación Oral •Lectura •Escritura
Habilidades	Habilidades menos integradas. Mayor énfasis en habilidades básicas (conocer y comprender)	Habilidades integradas y mayor énfasis en habilidades superiores. Aplicar, interpretar y analizar y evaluar, desarrollo de capacidad crítica y creativa.
Contenidos y énfasis	Hay énfasis en la lingüística y en los medios de comunicación por sobre la lectura y la escritura en reconocer estructuras gramaticales, ortográficos toda la E media Identificar elementos del lenguaje. Tipos de mensajes Elementos de la eficacia comunicativa Elementos del mundo literario Elementos del discurso argumentativo etc	Énfasis en la Literatura y se enriquece el eje de Escritura, convenciones y tipos de textos tiene carácter instrumental .

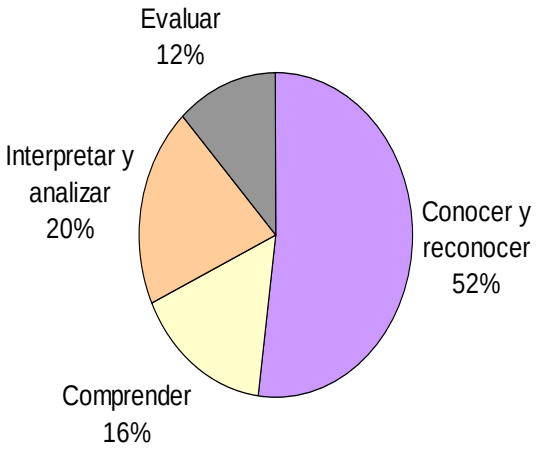
Lenguaje y Comunicación 2005 - 2009

III º medio 2005 OF	IIIº medio 2009 OF
1. **Comprender** los procesos de comunicación 2. Reconocer y utilizar los principales elementos, una evaluación crítica de la validez de los argumentos propios y ajenos…. 4. **Afianzar el dominio de las estructuras gramaticales** y léxico y la ortografía. 5. **Reconocer** la importancia para la cultura las obras literarias…. **y formarse una opinión personal** 6. **Comprender** y valorar la diversidad de visiones de mundo y de esas obras ofrecen 7. **Conocer** el contexto histórico cultural de la época en que se producen las obras leídas….. 8. **Reconocer** tanto la permanencia y transformaciones de elementos temático **identificando** los rasgos distintivos de las principales épocas y períodos que se distinguen en el proceso histórico de la literatura. 9. **Crear textos literarios** y no literarios que incorporen recursos y elementos del discurso argumentativo. 10. **Analizar críticamente** el discurso argumentativo en diferentes medios de comunicación escrita y audiovisual….. 11. **Reflexionar y tomar conciencia** del papel y responsabilidad de los medios de comunicación	1. **Interactuar** con propiedad en diversas situaciones comunicativas, predominantemente argumentativas….. 2. **Valorar la** comunicación verbal, no verbal y paraverbal al sustentar una posición 3. **Producir textos orales** de intención literaria y no literarios, 4. **Disfrutar la lectura de** obras literarias significativas, de distintos géneros, épocas y culturas, … 5. **Valorar con actitud crítica la lectura** de obras literarias,. 6. **Leer comprensivamente, con distintos** propósitos, textos en soportes impresos y electrónicos…. 7. **Leer comprensivamente variados tex**tos, identificando argumentaciones formadas por tesis y argumentos y evaluando la validez de los argumentos 8. **Leer comprensivamente, interpretando** y reinterpretando los sentidos globales de los textos…. 9. **Interpretar en los mensajes de los medios de comunicación** „ 10. **Producir, en forma manuscrita** y digital, textos de intención literaria y no literarios 11. **Utilizar adecuadamente un léxico** amplio y variado,. 12. **Escribir, utilizando flexible** y creativamente, de acuerdo con la estructura del texto. 13. Utilizar flexiblemente diferentes estrategias de escritura, 14. Valorar la escritura como una actividad creativa,

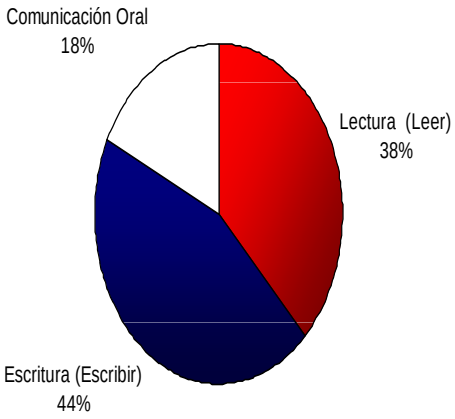
Lenguaje de comunicación-Ejes
Marco curricular 2005



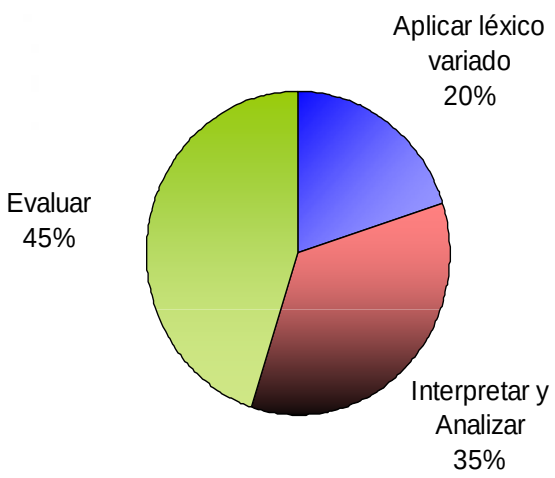
Lenguaje y Comunicación Marco curricular 2005
Habilidades Taxonomía Bloom



Lenguaje y Comunicación Ejes
Ajuste 2009



Lenguaje y comunicación
Ajuste 2009



Propuesta

- Centrar la prueba en comprensión de lectura.
- Eliminar las preguntas de contenido.
- Avanzar preguntas de redacción (ej: SAT)

Matemática

	2009
Ejes	Se organizó en torno a cuatro ejes comunes para básica y media ; •Números, Algebra, Geometría , Datos y azar considerando el razonamiento matemático transversal.
Contenidos	•Se eliminan vacíos y redundancias del currículum actual para mejorar la secuencia. •Números. Se posterga Nº irracionales para IIº medio •Funciones: se adelanta a 8º funciones •Algebra se introduce desde 5º básico •Datos y azar : se introduce en E media el modelamiento de la in certeza y se incrementan los contenidos de estadística.
Énfasis	•Se adelantan o agregan contenidos, y algunos tópicos se extienden en varios niveles, para alinear el currículum a marcos de pruebas internacionales •Se explicitan objetivos y contenidos relacionados con razonamiento matemático.

CIENCIAS

	2009
Ejes	•El currículum se organiza de 1° básico a 4° medio a partir de 5 ejes Estructura y función de los seres vivos; Organismos, ambiente y sus interacciones; Materia, energía y sus transformaciones; Fuerza y movimiento; La Tierra y el universo. •Se separa ciencias naturales y ciencias sociales en primer ciclo básico.
Contenidos	**CIENCIAS NATURALES Y BIOLOGÏA** •"Nutrición""digestión, circulación, respiración y excreción" exceptuando el tema de "enzimas" (1°), son tratados en el ajuste curricular en los niveles 5°, 6° y 8° año básico. •"Ciclos" (nitrógeno y carbono) se abordan en el ajuste en 7° año básico. •"Sistema muscular y respuesta motora" (3º), no está presente en el ajuste curricular. •Adaptaciónes" no está presente en el ajuste curricular. •**QUÍMICA** •"El Agua", "El Aire", "Los Suelos", "Los Materiales" han sido desglosados, en su mayor parte, a lo largo de la Enseñanza Básica. •"Los procesos químicos" no aparecen en el Ajuste Curricular. **FÎSICA** •El eje temático "Mundo atómico" no se trabaja en este nivel, ya que se trabaja en 1er año medio en el subsector Química.

Historia, Geografía y Ciencias Sociales

	2005	2009
Básica 5º a 8º		Espacio geográfico nacional continental y mundial y Geografía física •America pre colombina . • La conquista española •Instituciones de gobierno regional culturales económicas ,sociales no gubernamentales •La diversidad de las civilizaciones La herencia clásica: Grecia y Roma •La Europa medieval y el cristianismo •El humanismo y el desarrollo del pensamiento científico •La era de las revoluciones y de la conformación del mundo moderno •America Latina Contemporánea
Iº	1) Entorno natural y comunidad regional: geografía física y humana, y economía, problemas ambientales y diversidad cultural de las regiones de Chile. 2) Organización Regional. 3) Institucionalidad política: gobierno regional; poderes públicos, participación; conceptos políticos. 4) Sistema económico nacional: geografía económica de Chile; sistema económico del país.	•Primera Guerra Mundial y el mundo de entreguerras, Segunda Guerra Mundial, Guerra Fría: el mundo en la segunda mitad del siglo XX Geografía de la población Mundial •Proceso de urbanización en el siglo XX •El mundo globalizado con sus logros y falencias: internacionalización de la economía, organismos internacionales, TIC's, efectos de la globalización.

Historia, Geografía y Ciencias Sociales

IIº	Historia de Chile desde América precolombina hasta la actualidad	Chile, desde los pueblos originarios hasta el siglo XIX
IIIº	Historia universal, desde las grandes civilizaciones hasta la conformación del mundo contemporáneo (guerras mundiales).	Chile en el siglo XX, desde la crisis del parlamentarismo hasta la actualidad.
IVº	•El mundo contemporáneo •Orden mundial post-guerra (guerra fría) y orden mundial actual. •América Latina contemporánea. •4) Chile en el mundo actual: relaciones exteriores, económicas, tratados internacionales	•Estado de derecho en Chile (Constitución, institucionalidad, separación de poderes, sistema judicial). •Ejercicio de la ciudadanía, responsabilidades ciudadanas, desafíos de insertarse en un mundo globalizado. •Desafíos de las regiones de Chile. •Mercado del trabajo y legislación laboral

2010-2014

Niveles	2010	2011	2012	2013	2014
II° medio	Ajuste con texto ajustado sin programa	Ajuste con texto Sin programa	Ajuste con texto Sin programa		
I ° medio	Ajuste con texto ajustado Sin programa	Ajuste con texto ajustado Sin programa	Ajuste con texto Sin programa	Ajuste con texto Sin programa	
8° Bàsico	Ajuste con texto ajustado Sin programas	Ajuste con texto ajustado Con programas	Ajuste con texto ajustado Con programa	Bases	Bases
PSU			PSU con intersección de Ajuste 2009 / marco 2005	PSU con Intersección ajuste2009 / marco 2005	PSU con Intersección Ajuste /bases

Conclusiones

- Cambios mayores en ajuste 2009 respecto al 2005 son en Lenguaje ,Historia y Ciencia .
- Generación que rendirá PSU 2012 re quiere alineamiento con ajuste por textos
- Importancia de alineamiento con PSU para curriculum.
- Importancia de un mensaje tranquilizador y clarificador con información oportuna.

Appendix AA. *Comité Técnico Asesor*

On the following pages, the evaluation team has inserted a document provided by MINEDUC concerning the *Comité Técnico Asesor del Consejo Directivo para Las Pruebas De Selección y Actividades de Admisión: Operacionalización de Funciones y Atribuciones.* This attachment fully articulates the roles and responsibilities of the Comité Técnico Asesor (CTA, or Technical Advisory Committee, in English).

"COMITÉ TÉCNICO ASESOR DEL CONSEJO DIRECTIVO PARA LAS PRUEBAS DE SELECCIÓN Y ACTIVIDADES DE ADMISIÓN"

Operacionalización de funciones y atribuciones

I. De la naturaleza y conformación del Comité Técnico Asesor (CTA)

Art. 1. El CTA es un organismo del Consejo de Rectores de las Universidades Chilenas (**CRUCH**) cuya misión es *colaborar con el Consejo Directivo para las Pruebas de Selección y Actividades de Admisión (**CD**) en aquellas tareas que éste le encomiende como parte de su función general de "coordinación y supervisión de la institucionalidad que rige el conjunto de las dimensiones de selección y admisión a las universidades del Consejo de Rectores",* además es *"la entidad intermediadora entre el CD y los equipos técnicos responsables de la elaboración y aplicación de las pruebas* de selección" (PSU).

Art. 2. El CD es el organismo del CRUCH que tiene a su cargo los temas de selección y admisión de alumnos en el marco de las políticas que el CRUCH determine al respecto. En este sentido, tiene tuición sobre los aspectos técnicos del proceso, en especial en lo referente al DEMRE, según se detalla más adelante.

Art. 3. El CTA estará formado por un grupo de académicos provenientes de universidades que conforman el CRUCH[1]. Su número, composición y tiempo de ejercicio, así como su designación serán decididos por el CD, el que también designará al Presidente del CTA.

Art. 4. El CTA se reporta directamente al CD y al Vicepresidente Ejecutivo del CRUCH.

II. Funciones.

Art. 5. El CTA tiene funciones generales, de difusión, de supervisión, de coordinación, de estudios, y ejecutivas.

[1] Las universidades del CRUCH se comprometen a facilitar la participación de sus académicos en el CTA así como a darles las facilidades para cumplir sus funciones, reconociendo esta labor como dedicación académica.

II.1. Funciones generales.

Art. 6. Corresponde al CTA, en el marco de su misión:

a. Conocer de las tareas encomendadas por el CD en tiempos y plazos oportunos;

b. Solicitar al CD los recursos necesarios para llevar adelante los trabajos cuando éstos no se encuentren programados y presupuestados;

c. Acudir al CD para zanjar cuestiones que excedan su capacidad;

d. Proponer iniciativas al CD vinculadas con aspectos propios de su función técnica (tratamiento de datos, ponderaciones, conversiones, etc) así como relacionados con otros aspectos tales como políticas de selección, vinculación con organismos internacionales, entre otros;

e. Informar periódica y sistemáticamente, al menos dos veces al año, al CD de los resultados de su trabajo y de las tareas encomendadas. Al inicio de cada año académico, presentar al CD una agenda de trabajo para el año en curso, la que deberá ser aprobada por el CD e incluida en la pauta de financiamiento;

f. Dar cuenta del uso de los recursos que le han sido asignados para su funcionamiento por el CD; y

g. Proponer soluciones técnicamente apropiadas sobre aquellas cuestiones que le encomiende el CD[2].

II.2. Función de difusión.

Art.7. Corresponde al CTA c*oordinar y supervisar el programa de difusión e información masiva sobre las Pruebas de Selección y Actividades de Admisión dirigido al Magisterio y estudiantes de la Enseñanza Media*.

Art.8. Para ello deberá:

[2] Por ejemplo, durante 2004 se recibieron consultas relativas a la equivalencia del Baccalauréat francés y el Abitur alemán. Entre otros temas, la Prueba especial de Tecnología para egresados de la ETP; incorporación de prueba del idioma inglés; administración de las pruebas más de una vez al año.

a. Con antelación al inicio del proceso de admisión de cada año, coordinar con el DEMRE la estrategia de difusión, y establecer la programación y contenidos de la información asociada al proceso de selección que será entregada a los usuarios del sistema;

b. Proponer contenidos específicos para entregar a los postulantes y público en general, vía suplementos del diario que sea el órgano oficial de difusión del proceso, relacionados con los tópicos técnicos que corresponden al CTA;

c. Producir la información técnica apropiada para ser difundida masivamente por la prensa y en las páginas y portales web asociados (CRUCH, DEMRE, Mineduc, universidades, otros); y

d. Organizar y celebrar, a lo menos, un seminario de difusión sobre aspectos técnicos del proceso, destinado a especialistas y estudiosos del área.

II.3. Función de supervisión.

Art. 9. Corresponde al CTA *supervisar y conocer de los procesos de elaboración, experimentación, aplicación, procesamiento y entrega de los resultados de las pruebas*.

Art. 10. Para ello deberá:

a. Solicitar y recibir del DEMRE la información necesaria acerca de los procesos atingentes a las pruebas experimentales y definitivas en su planificación, elaboración, aplicación y análisis de resultados;

b. Supervisar los procesos analíticos de las pruebas y sus resultados, con asistencia personal de miembros del CTA a las dependencias del DEMRE durante el procesamiento de los datos, antes de su entrega a los postulantes;

c. Colaborar con el DEMRE en la solución de problemas técnicos relacionados con la construcción de las pruebas, su aplicación y procesamiento de los resultados de las mismas;

d. Colaborar en decisiones sobre ajustes y linking en la prueba de ciencias u otras situaciones que lo demanden, durante la asistencia personal de miembros del CTA a las dependencias DEMRE durante el procesamiento de los datos;

e. Tener iniciativa para proponer al CD modificaciones a los procesos métricos de las pruebas en orden a mejorar la calidad, confiabilidad y transparencia de los procesos y sus resultados; y

f. Preparar informes técnicos pertinentes para el CD, DEMRE y archivo del CTA, indicando la racionalidad que sustentan las observaciones y/o propuestas que se hagan en cada caso. Estos informes deben estar disponibles para el público especializado que desee consultarlos.

II.4. Función de coordinación.

Art. 11. Corresponde al CTA coordinar y supervisar la marcha de los procesos antes señalados (requeriendo y recibiendo información oportuna y adecuada de los equipos técnicos responsables) y reportar al Consejo Directivo (y a través de éste al Consejo de Rectores) con la frecuencia que éste determine.

Art. 12. Para ello deberá:

a. Reunirse al menos 4 veces por año con el DEMRE (Director y personal superior); en dichas instancias se abordarán de manera principal los siguientes tópicos, documentando en acta los acuerdos que se tomen al respecto:

1. Planificación de los procesos anuales: pruebas experimentales, pruebas definitivas, difusión.

2. Conocimiento y análisis de los resultados de las pruebas experimentales; impacto sobre el banco de ítemes de las pruebas.

3. Revisión de los procesos analíticos y de procesamiento de datos para las pruebas anuales; decisiones sobre ajustes y linking (caso de ciencias)

4. Conocimiento y análisis de los resultados de la aplicación anual de las pruebas PSU. Planteamiento de los requerimientos de apoyo.

Evaluación de la viabilidad de equipos y software para los grados de complejidad creciente del proceso.

b. Colaborar con el DEMRE en la optimización de los procesos anuales, en el examen de los resultados experimentales, en el procesamiento de los datos de las pruebas, y en el análisis de los resultados del proceso de medición y selección en sus diversos aspectos;

c. Reportar, al menos bianualmente, al CD los avances y resultados de la gestión del CTA, según plan de trabajo aprobado; y

d. Solicitar - anualmente - a los Registradores Curriculares de las Universidades miembros del H. Consejo de Rectores (o quienes hayan sido encomendados) la información relativa a los rendimientos de los estudiantes del primer año de las diferentes carreras de cada universidad, con el propósito de realizar los estudios de Validez Predictiva correspondientes.

II.5. Función de estudios.

Art. 13. Corresponde al CTA la realización, directa o indirecta, de estudios científicos en materias de medición y selección, que profundicen, entre otros aspectos, en el conocimiento de la naturaleza y características sociales y académicas de los postulantes, de las capacidades de los sistemas de puntajes, de los niveles de logro curricular de los postulantes.

Art. 14. Para ello el CTA deberá:

a. Producir y difundir, anualmente, un informe en profundidad acerca de las características que ha revestido la aplicación así como los resultados de las PSU y los puntajes de las NEM;

b. Producir y difundir, anualmente, un informe acerca de las características métricas de los instrumentos utilizados en el proceso inmediatamente anterior, y su comparación con características de otras aplicaciones así como con las propiedades de instrumentos de reconocida calidad utilizados en otros países;

c. Producir y difundir, anualmente, los informes relativos al potencial predictivo de las series de puntajes utilizadas en la selección, así como la consideración de otras variables asociadas al fenómeno del

rendimiento estudiantil, particularmente en el primer año de estudios universitarios; y

d. Estudiar e informar, anualmente, la situación de los estudiantes ingresados al sistema universitario por vías diferentes a la establecida, en particular cuando se trata de exámenes válidos para la admisión en otros países

II.6. Función ejecutiva.

Art. 15. Corresponde al CTA la dirección de los procesos relacionados con las decisiones y operaciones que tienen que ver con cuestiones propiamente técnicas de la instrumentación, procesamiento y análisis de datos, del proceso general de selección de alumnos a las universidades.

Art. 16. Para ello, el CTA deberá:

a. Concordar anualmente con el DEMRE las características técnicas que deben ser observadas en los procesos de instrumentación, procesamiento y análisis de los datos de las pruebas de selección a la universidad; y

b. Señalar al DEMRE aquellos aspectos que ameritan especial consideración en materia de difusión a la comunidad, a los estudiantes, a los colegios y profesores, así como a las autoridades de educación y universitarias.

III. Del Presidente.

Art. 17. El Presidente del CTA es un académico miembro de dicho comité que ha sido designado en tal calidad por el CD por el período que este último determine.

Art. 18. Son atribuciones del Presidente:

a. Representar al CTA ante el CD y las autoridades educacionales;

b. Actuar como vocero del CTA en las materias que le competen directamente;

c. Presidir las sesiones ordinarias y extraordinarias del CTA;

d. Citar al CTA de manera extraordinaria;

e. Citar al Director del DEMRE a reuniones de trabajo; y

f. Solicitar al DEMRE y a las universidades del CRUCH, aquella información que sea necesaria para el quehacer del CTA en las funciones que le son propias.

Art. 19. Son deberes del Presidente:

a. Responder ante el CD de los actos y decisiones del CTA; y

b. Autorizar los gastos del CTA.
