

PRUEBA INICIA

**EDUCATION GLOBAL PRACTICE
LATIN AMERICA AND CARRIBEAN REGION
THE WORLD BANK**



August 2014

INTRODUCTION.....	1
IMPROVING THE QUALITY OF EDUCATION	1
LOW LEARNING LEVELS	1
GETTING HIGH QUALITY TEACHERS IN ALL CLASSROOMS	2
WHAT IS THIS REPORT ABOUT AND HOW DID IT COME ABOUT?	3
HOW IS THIS REPORT ORGANIZED?.....	6
IDENTIFYING GOOD TEACHERS	7
DETERMINANTS OF TEACHER EFFECTIVENESS	8
<i>Teacher Education</i>	11
<i>Experience</i>	13
<i>Content Versus Pedagogical Knowledge</i>	13
<i>Teacher Certification and Licensing</i>	13
<i>Teacher Professional Development (TPD)</i>	14
THE PRUEBA INICIA AND ITS PROPERTIES.....	15
OBJECTIVES OR PURPOSE OF THE TEST	15
THE DEVELOPMENT PROCESS.....	16
STANDARD SETTING	18
SUMMARY REVIEW OF A SAMPLE TEST (EXAMPLE: MATHEMATICS).....	20
<i>Ownership</i>	20
<i>Purpose of the Test</i>	20
<i>Table of Test Specification</i>	20
<i>Item Development and Tasks</i>	21
<i>Piloting</i>	22
<i>Validation Concerns</i>	23
<i>Documentation</i>	25
PSYCHOMETRIC PROPERTIES	27
<i>Item Analysis</i>	27
WHAT DO TEACHERS HAVE TO SAY ABOUT THE PRUEBA INICIA	29
THREE KEY CONCLUDING POINTS	30
BENCHMARKING THE PRUEBA INICIA.....	31
QUALITY ASSURANCE MECHANISMS	32
TEACHER INTAKE	32
EXIT REQUIREMENTS	34
<i>Licensing and Certification</i>	34
<i>Importance of Teaching Practice</i>	35
<i>Coaching and Mentoring</i>	36
<i>Performance Based Assessment</i>	37
<i>Curricular Design</i>	37
CONCLUSIONS AND POLICY OPTIONS.....	39
DETAILED PSYCHOMETRIC ASSESSMENT	42
METHODOLOGICAL ISSUES	43

ANALYSIS	43
PSYCHOMETRIC CONCEPTS AND METHODS OF ANALYSES	43
<i>IRT-Modelling and Classical Test Theory</i>	43
<i>Specifics of IRT/Rasch Modelling</i>	44
<i>Specifics of Classical Test Theory and Related Indicators</i>	45
PSYCHOMETRIC ANALYSES	47
INTRODUCTION	47
RESTRICTIONS OF THE ANALYSIS	48
RESULTS	48
<i>Development and Implementation of the INICIA</i>	48
<i>Face Validity</i>	50
<i>Construct Validity</i>	50
<i>Content Validity</i>	51
<i>Ecological Validity</i>	52
PSYCHOMETRIC CHARACTERISTICS OF THE TESTS	53
OVERALL QUALITY OF THE TESTS AND ITEMS	53
DISTRACTOR ANALYSIS OF THE ITEMS	60
<i>Summary of Distractor Analysis</i>	75
DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSIS	75
ITEM PARAMETERS FROM THE IRT MODELLING	75
EQUATED SCORES OVER TESTS	76
ADEQUACY AND COMPARABILITY OF THE REPORTING CATEGORIES	77
CONCLUSIONS AND SUMMARY	80
REFERENCES	84
APPENDIX A	88
ITEM PARAMETERS OF THE ITEMS IN INICÍA	88
<i>Table A.1A Item parameters of PCD-Básica</i>	89
<i>Table A.2A Item parameters of PCP-Básica</i>	93
<i>Table A.3 Item parameters of PCD-Biología</i>	96
<i>Table A.4 Item parameters of PCD-Física</i>	98
<i>Table A.5 Item parameters of PCD-Matemática</i>	100
<i>Table A.6 Item parameters of PCD-Química</i>	102
<i>Table A.7 A Item parameters of PCD-Historia</i>	104
<i>Table A.8 A Item parameters of PCD-Lenguaje</i>	108
<i>Table A.9A Item parameters of PCD-Parvularia</i>	111
<i>Table A.10A Item parameters of PCP-Parvularia</i>	114
<i>Table A.11A Item parameters of PCP-Media</i>	117
APPENDIX B	120
CHARACTERISTICS OF THE FLAGGED ITEMS IN INICÍA	120
<i>Table B.1A: Poor or pathological items in PCD-Básica</i>	121
<i>Table B.1B: Poor or pathological items in PCD-Básica Version B</i>	126

<i>Table B.2A Poor or pathological items in PCP-Básica Version A</i>	131
<i>Table B.2B Poor or pathological items in PCP-Básica Version B</i>	134
<i>Table B.3 Poor or pathological items in PCD-Biología</i>	137
<i>Table B.4 Poor or pathological items in PCD-Física</i>	141
<i>Table B.5 Poor or pathological items in PCD-Matemática</i>	143
<i>Table B.6 Poor or pathological items in PCP-Química</i>	146
<i>Table B.7A Poor or pathological items in PCD-Historia Version A</i>	151
<i>Table B.7B Poor or pathological items in PCD-Historia Version B</i>	155
<i>Table B.8A Poor or pathological items in PCD-Lenguaje Version A</i>	159
<i>Table B.8B Poor or pathological items in PCD-Lenguaje Version B</i>	164
<i>Table B.9A Poor or pathological items in PCD-Parvularia Version A</i>	170
<i>Table B.9B Poor or pathological items in PCD-Parvularia Version B</i>	176
<i>Table B.10A Poor or pathological items in PCP-Parvularia Version A</i>	181
<i>Table B.10B Poor or pathological items in PCP-Parvularia Version B</i>	184
<i>Table B.11A Poor or pathological items in PCP-Media Version A</i>	187
<i>Table B.11B Poor or pathological items in PCP-Media Version B</i>	191
APPENDIX C	193
DIF ANALYSIS OF THE ITEMS IN INICÍA	193
<i>Table C.1A DIF of PCD-Básica Version A</i>	194
<i>Table C.1B DIF of PCD-Básica Version B</i>	197
<i>Table C.2A DIF of PCP Básica Version A</i>	199
<i>Table C.2B DIF of PCP-Básica Version B</i>	202
<i>Table C.3 DIF of PCD-Biología</i>	204
<i>Table C.4 DIF of PCD-Física</i>	207
<i>Table C.5 DIF of PCD-Matemática</i>	210
<i>Table C.6 DIF of PCD-Química</i>	212
<i>Table C.7A DIF of PCD-Historia Version A</i>	214
<i>Table C.7B DIF of PCD-Historia Version B</i>	216
<i>Table C.8A DIF of PCD-Lenguaje Version A</i>	218
<i>Table C.8B DIF of PCD-Lenguaje Version B</i>	220
<i>Table C.9A DIF of PCD-Parvularia Version A</i>	221
<i>Table C.9B DIF of PCD-Parvularia Version B</i>	223
<i>Table C.10A DIF of PCP-Parvularia Version A</i>	225
<i>Table C.10B DIF of PCP-Parvularia Version B</i>	227
<i>Table C.11a DIF of PCP-Media Version A</i>	229
<i>Table C.11B DIF of PCP-Media Version B</i>	231
APPENDIX D	234
EQUATED SCORES IN INICÍA	234
<i>Table D.1 Equated scores in PCE-INICÍA</i>	235
<i>Table D.3 Equated scores in PCP-Básica Versions A and B</i>	241
<i>Table D.4 Equated scores in PCD-Biología</i>	244
<i>Table D.5 Equated scores in PCD-Física</i>	246
<i>Table D.6 Equated scores in PCD-Matemática</i>	248

<i>Table D.7 Equated scores in PCD-Química.....</i>	<i>250</i>
<i>Table D.8 Equated scores in PCD-Historia Versions A and B.....</i>	<i>252</i>
<i>Table D.9 Equated scores in PCD-Lenguaje Versions A and B.....</i>	<i>256</i>
<i>Table D.10 Equated scores in PCD-Parvularia Versions A and B.....</i>	<i>260</i>
<i>Table D.11 Equated scores in PCP-Parvularia Versions A and B.....</i>	<i>264</i>
<i>Table D.12 Equated scores in PCP-Media Versions A and B.....</i>	<i>267</i>

List of Boxes

Box 1: Weak Quality Assurance in Higher Education.....	5
Box 2: The Problem of Identifying Good Teachers Using Ex-Ante Measures	10
Box 3: NCTQ Study on Quality of Teacher Preparation in the United States.....	11
Box 4: Standards for the NCTQ Teacher Prep Review	12
Box 5: Building a Better Teacher: How Teaching Works (and How to Teach It to Everyone)	14
Box 6: Un Buen Comienzo	36

List of Figures

Figure 1 : Number of Teacher Training Institutions in Chile (1980-2008)	3
Figure 2 : Does Training Matter	9
Figure 3 : General Results – Prueba Inicia 2012	19
Figure 4 : Content Knowledge – Secondary – Prueba Inicia 2012	20
Figure 5 : Good Practice on Textbook Design	21
Figure 6 : Mathematics Standards Table for Secondary Schools	25
Figure 7 : Comparing The Praxis and Prueba Mathematics Examination	26
Figure 8 : Mechanisms for Quality Assurance	32
Figure 9 : Relationship between item difficulty and item discrimination in the <i>PCE-INICÍA</i>	54
Figure 10 : Item Discrimination and Difficulty of <i>PCD-Básica</i>	55
Figure 11 : Item Discrimination and Difficulty of <i>PCD-Básica</i> and <i>PCP-Básica</i>	55
Figure 12 : Item discrimination and -difficulty of <i>PCD-Matemática</i>	56
Figure 13 : Item discrimination and -difficulty of <i>PCD-Física</i>	56
Figure 14 : Item discrimination and -difficulty of <i>PCD-Biológica</i>	57
Figure 15 : Item discrimination and -difficulty of <i>PCD-Química</i>	57
Figure 16 : Item discrimination and -difficulty of <i>PCD-Historia</i>	58
Figure 17 : Item discrimination and -difficulty of <i>PCD-Lenguaje</i>	58
Figure 18 : Item discrimination and -difficulty of <i>PCD-Parvularia</i>	59
Figure 19 : Item discrimination and -difficulty of <i>PCP-Parvularia</i>	59
Figure 20 : Item discrimination and -difficulty of <i>PCP-Media</i>	60
Figure 21 : An example of a graphical distractor-wise analysis of a good item	61
Figure 22 : An example of a distractor-wise analysis of a pathological item: a possible wrong key	62
Figure 23 : An example of a distractor-wise analysis of a pathological item: an item with no alternative for the correct answer	62
Figure 24 : An example of a distractor-wise analysis of a pathological item: Pathological Guessing	63
Figure 25 : An example of a distractor-wise analysis of a pathological item: Several Correct Answers ...	63
Figure 26 : A selection of suspicious or pathological items in <i>PCD-Básica</i>	64
Figure 27 : A selection of suspicious or pathological items in <i>PCP-Básica</i>	65
Figure 28 : A selection of suspicious or pathological items in <i>PCD-Biológica</i>	66
Figure 29 : A selection of suspicious or pathological items in <i>PCD-Física</i>	67
Figure 30 : A selection of suspicious or pathological items in <i>PCD-Matemática</i>	67
Figure 31 : Selection of suspicious or pathological items in <i>PCD-Química</i>	69
Figure 32 : Selection of suspicious or pathological items in <i>PCD-Historia</i>	70
Figure 33 : Selection of suspicious or pathological items in <i>PCD-Lenguaje</i>	71
Figure 34 : Selection of suspicious or pathological items in <i>PCD-Parvularia</i>	72
Figure 35 : Selection of suspicious or pathological items in <i>PCP-Parvularia</i>	73
Figure 36 : A Selection of suspicious or pathological items in <i>PCP-Media</i>	74

Acknowledgements

Leave this page blank

INTRODUCTION

Improving the Quality of Education

1. **To say that Chileans are passionate about the educational opportunities available in the country would be an understatement.** The debate over education is a daily affair in the country and the discussion is everywhere – in print, television news, blogging sites, and every so often, it spills over into the streets. On August 22, 2014, students numbering in the thousands were out on the streets of Santiago and elsewhere in the country, demanding free education of high quality across all levels. Almost exactly four years ago, during the previous government’s tenure, students had poured out on to the streets of Santiago making an almost identical set of demands. Ironically, across a span of four years and under two political parties representing opposite sides of the political spectrum, an identical set of instruments were used to deal with the protestors – water cannons, tear gas, and mass arrests.

2. **Over the last forty years, perhaps no other sector in Chile has witnessed the monumental shifts in government policy as has the education sector.** Chile’s military government introduced sweeping changes that completely altered the administrative, financial and delivery models for education in the country. They ushered in market-oriented mechanisms, decentralized school administration, introduced incentives to support the expansion of state financed private schooling, and dramatically altered the status of teachers – eliminating their positions as civil servants. With the return to democracy in 1990, education policy has taken a renewed focus with equity and quality as its central objectives, with an emphasis on student learning, and a teacher management policy where teachers once again have tenured assignments.

3. **However, in these intervening years, the nature of the problem has changed.** The early reforms were aimed at getting children into schools and the use of public financing, coupled with private management, helped in dramatically alleviating supply side constraints and expanding access to educational opportunities. Chile has been very successful in this regard. The more recent reforms are aimed at improving the quality of education, and in ensuring that all children have access to such opportunities. These reforms are complex and stated targets cannot be met through a simple expansion of resources flowing into these programs. A comprehensive and holistic view of the problem is needed and the tools to address the existing constraints need to be developed with these in mind. Not only is it important to have a deeper understanding of the various inputs that would be needed, but it is also important to ensure an understanding of *how* these inputs will be brought into play. The issues are made more complex because while there is a broad agreement across stakeholders on some goals – for example, the need to improve educational quality - how and in what manner this should be done has much less consensus. For example, the current set of protests by students’ focuses on the role of the students, by all accounts an important stakeholder in the education process, in the governance structures of educational institutions. As a key stakeholder in the process, students believe that they should have a seat at the policy table to ensure that their voices are heard and that they can influence decision making in the sector. Other stakeholders do not necessarily agree. Notwithstanding these debates, there is general consensus in the country that the quality of education needs to be improved.

Low Learning Levels

4. **In the Latin America and Caribbean (LAC) region, Chile is considered to be a star performer in the education sector.** It consistently sits atop the continental leaderboard on numerous

education indicators. Chile's educational successes have become a model for many other countries in the region and beyond¹.

5. **However, there remain serious concerns regarding education quality and how quality education can be made available to all students.** Though a regional powerhouse, Chile performs poorly when compared to the best in the world, and falls well short of OECD averages in terms of achievement scores in global standardized assessments. For example, in the past three PISA assessments, Chile has found itself clustered towards the lower end of the performance distribution. In the most recent round, carried out across 65 participating countries in 2012, Chile ranked 51 (in Math), 46 (in Science) and 47 (in Reading)².

6. **Though the PISA 2012 results were released after the conclusion of the recently held presidential elections in November 2013 in Chile, policymakers were compelled to respond to the country's relatively poor showing and law-makers from across the political spectrum pledged to work on improving school quality and increasing learning outcomes.** This sort of debate is to be expected in Chile. Having become a member of the OECD, the expectations of the Chilean population are set even higher. Consequently, there is a strong demand for improved education quality across all levels, with the population at large and the Government not happy with the fact that the country is punching well below where its economic weight would predict it should.

Getting High Quality Teachers in all Classrooms

7. **Improving the teacher quality and ensuring that only qualified teachers are placed in classrooms has become a fundamental Government priority.** Schooling quality is determined by many factors, both teacher and non-teacher factors³. The latter includes *inter alia* student motivation and incentives, infrastructure, technology, expenditure per pupil, curriculum, etc. Likewise teacher related factors are numerous – including selection and training in teacher education programs, the teacher recruitment process, compensation, incentives and career ladder, and continuous professional development and training. This latter set of issues which are centered around the teacher are particularly important as recent research illustrates that the quality of the teacher in the classroom is perhaps the single most important factor affecting student learning outcomes. Thus, a *strong teaching force* is essential to improving the quality of the schooling experience and school education. High quality teaching forces are a common feature of countries with high quality educational systems. Teacher related factors are also important for another reason. It is a lever that can be manipulated relatively easily through public policy, which is not true for parental background, gender, socio-economic status, etc. While this much is known, how to develop, recruit, deploy, motivate and compensate teachers to do their tasks every day in the classroom is still not clear despite innumerable studies, and efforts to determine what makes a *teacher great*.

8. **The Chilean government is keenly aware of the paramount importance of having a high quality and effective teacher in every class.** Teacher-related factors, especially, the formative work needed to prepare a teacher prior to placing them in front of students has become an important area both for research and policy formulation. This search for a solution to address concerns of quality has led to a

¹ These influential policies are not restricted to those that have been in place since 1990, but include those that were put in place by the far right governments before – the role of the private sector, the use of vouchers, results-based financing for tertiary education institutions, teacher policies, and program evaluation.

² In the TIMSS round of 1999, out of 38 participating nations, Chile finished fourth from the bottom in both Math and Science assessments for students in Grade 8.

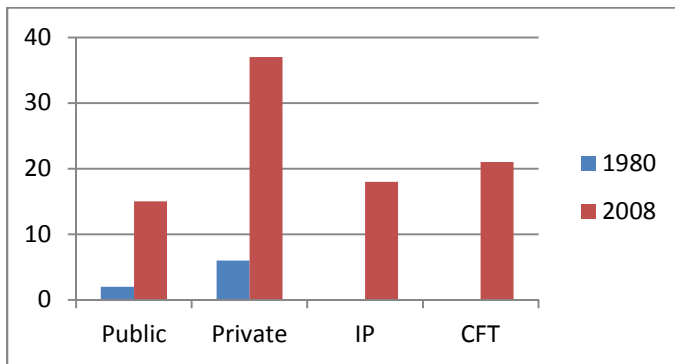
³ Of course, it is also based on numerous non-school factors, such as, parental education, socio-economic status, gender, etc.

considerable shift in the teacher preparation paradigm, and as Meckes et al. (2012) state “...a transition from policies that provided support to initial teacher training improvement initiatives with low-stakes accountability measures in the late nineties, to policies that combine support, incentives, pressure and high-stakes accountability.”

9. **This focus on high stakes accountability measures has continued and the set of tools to ensure accountability has grown.** Koljatic and Silva (2013) illustrate the increased use of measurement and assessments tools in the Chilean context and across the education spectrum from admissions to pre-kindergarten to selection for universities. They also go on to illustrate that this growth in the use of assessments is happening but with a limited understanding of the technical requirements needed for the use of test scores and evaluations of this nature. The authors show that not only are some of the assessments poorly developed, but many are also used inappropriately. They call for improved and well developed policies and guidelines for the use and deployment of such assessments.

10. **The number of teacher preparation institutions and centers in Chile has increased dramatically between 1980 and today.** In addition to the shifting accountability paradigm, there are other reasons why there should be concerns regarding teacher preparation in Chile. In particular, the dramatic growth in the number of institutions and centers that are involved in teacher education and preparation in the country. The figure below illustrates this dramatic growth between 1980 and 2008. This dramatic increase, fueled mostly by growth in private sector institutions, coupled with the fact that the quality assurance systems for tertiary programs are still relatively weak, suggests that many teachers are probably entering classrooms ill-prepared both in terms of content knowledge and pedagogical ability.

Figure 1
Number of Teacher Training Institutions in Chile (1980-2008)⁴



What is this report about and how did it come about?

11. **It is with this concern over education quality that the Government of Chile has sought the assistance of the World Bank in reviewing one specific aspect of their teacher development system - the Prueba Inicia.** As part of its strategy to improve teaching quality, Chile introduced in 2008 a voluntary⁵ teacher assessment to monitor the knowledge and skills of new graduates emerging from pre-teacher training institutions. The teacher-trainee exit assessment is conducted annually and is increasingly gaining popularity amongst both teacher training institutions and the trainees themselves. At

⁴ Education Internacional Latin America Regional Office (2010)

⁵ Though there are calls to make this a mandatory step in the process of selecting and recruiting teachers.

present, the Inicia is a voluntary assessment⁶ and teacher trainees may opt out of being assessed. However, the policy focus is on developing this instrument further with the eventual aim of making it mandatory and as one element in the process of ensuring that all students have in their classrooms a highly qualified and effective teacher.

12. The combination of dramatic growth in student numbers and these weak schooling outcomes have necessitated a review of key processes. Between 1981 and 2012 the number of universities in Chile has grown from 8 to 60, and in the ten years between 2002 and 2012, the student population at the tertiary level has mushroomed from about 500,000 to almost 1,100,000. This impressive growth coupled with the fact that the quality control mechanisms were not fully developed and deployed, implies that across the tertiary space there is a need to ensure mechanisms for quality control. Furthermore, given that the share of students in teacher preparation is about 14 percent of the total, suggests that in particular, quality assurance of the graduates of teacher training programs is essential.

13. Across the world, it is possible to categorize two distinct paths for ensuring quality of teacher training graduates. These can be broadly defined as upstream and downstream measures.

- Upstream measures are those whereby institutional quality is ascertained through quality control mechanisms (for example, accreditation of teacher training institutions) which then are responsible for ensuring high quality graduates⁷.
- Downstream processes or filtering mechanisms are those whereby graduates of teacher training programs are required to pass some set of standardized assessments before obtaining a license⁸ to function as a teacher⁹. The specific requirement for licensing differs from country to country, and in federal structures like the United States, the requirements differ from state to state.

14. In Chile, the major concern is that both upstream and downstream processes are weak. Institutional accreditation and quality assurance mechanisms do exist at the tertiary level, but outside of

⁶ Given the voluntary nature of the assessment till date, one cannot even refer to it as a licensing exam, though for all practical purposes the Inicia is a licensing assessment. Presently, the results are released aggregated at the institutional level or higher. There are no consequences at the individual level for a poor performance at this point in time. It is a teacher trainee exit exam, aimed at measuring the skills gained by the trainee in teacher training institutions – both content and pedagogical - and not an instrument meant for selection. However, the policy debate around the INICIA almost treats it as a selection tool. For example, under the previous government there were plans to not only make the INICIA a mandatory assessment, but in part to link initial teacher compensation to their results in the INICIA. So, while the purists refer to the INICIA as a teacher trainee exit examination, largely assessing content knowledge, in the policy world it is seen as a much more potent tool for some. So, in this report, at times we may treat the INICIA as being more than an exit examination.

⁷ In this case, students merely have to complete the requirements of the institutions in which they are studying and this automatically qualifies them for entry into the profession. Many high performing countries use this approach with Singapore perhaps presenting an extreme approach with only a single teacher training institution that caters to all of the island's needs. Examples of countries which employ downstream processes included the US, UK and a range of other countries.

⁸ Similar licensing requirements exist in other professions as well and the processes entailed vary from country to country. For example, licensing is a common feature of the engineering profession in many countries and is typically done to ensure public safety and welfare, and other similar interests. This is also true for other professions where public safety and welfare have to be safeguarded by the state – such as, medicine, law, accounting, etc.

⁹ This is perhaps used more in developing country settings where guaranteeing institutional quality through upstream mechanisms is harder to undertake.

health and education, institutional accreditation is a voluntary mechanism. Licensing is mandatory and all institutions wishing to operate in the tertiary space need to be licensed (OECD 2012). However, there are inherent weaknesses in the system and these are highlighted in Box 1. On the downstream side, since 2008, the GoC has put in place the Prueba Inicia, a voluntary exit exam for teacher. Given the voluntary nature of the assessment, only about 3,200 students from 49 teacher training institutions participated in these assessments (or about 2.5% of the total number of students in teacher training programs¹⁰) in 2011. So controls on both sides are weak. Furthermore, of those who participated, about 69% demonstrated “insufficient” content knowledge in relevant subject areas; and in some institutions, more than 90% of their graduates obtained “insufficient” results.

Box 1

Weak Quality Assurance in Higher Education¹¹

- The report entitled *Quality Assurance in Higher Education in Chile* (OECD, 2012), states that *key quality assurance principles* are not fully addressed by the SINEACES or the quality assurance mechanism in Chile in a number of dimensions as shown below:
- Basic assurance of minimum standards is not consistently provided
 - A quality culture which embraces continuous improvement is still only emerging
 - The role of users – notably students and employers - in assuring quality is peripheral
 - The system has been developed with the missions, practices and aspirations of the longer-established universities in mind, and to be more suited to them. This tendency for a „one size fits all“ approach is perceived to be unsuited to the development of vocational and professionally oriented institutions
 - A lack of transparency about how decisions are made within SINAC-ES has weakened confidence within the system and created mistrust in the public mind about the judgments that it makes
 - The lack of an integrated and verifiable information system has led to a loss of trust in the data which is provided, and contributed to a situation in which information can be misleadingly presented
 - There appears to be no clear strategy for international engagement

15. The Government of Chile is eager to strengthen the country’s teacher preparation program given that many countries with high quality teaching systems have systems in place to ensure high quality teachers. A key feature of this effort focuses on strengthening the Prueba Inicia – in terms of its design, its implementation, and most importantly its coverage. The continued poor performance of teacher trainees in this assessment has triggered a rich debate about how to improve the quality of teacher training programs in Chile. This is compounded by concerns of whether the instrument itself is an appropriate one to measure the skills needed by teacher trainees as they transition to classroom teachers and are placed in schools across the country.

16. This is what has prompted the Government to seek the World Bank’s assistance. There are three key objectives of this exercise. These include: (i) Benchmarking the Prueba Inicia against similar practices in a select set of countries, (ii) a detailed psychometric assessment of the Prueba Inicia instrument, and (iii) a set of policy recommendations regarding the Prueba Inicia and its uses.

¹⁰ Captures the numbers of test takers against the *total number* of students enrolled in teacher preparation programs and not the total number of students in the *final year* of their programs. In 2012, the proportion of test takers in the total number of students exiting teacher training programs was about 14% (Ministry of Education, 2013).

¹¹ OECD Study on Quality Assurance in tertiary institutions in Chile

How is this report organized?

This report is organized into two parts.

Part I focuses on the three objectives mentioned above.

- It begins with a review of what makes a teacher good, why this is important, but also why it is difficult to identify a good teacher *ex ante*. While the notion of having a high quality teacher in every classroom is easily understood and intuitive, the challenges in this regard are far more subtle and not easily understood. This section will review the difficulties associated with identifying the characteristics of a good teacher and will explain why what we believe to be intuitive measures are not necessarily good predictors of performance in the classroom.
- The second section focuses on the Prueba Inicia, the standards on which it is based and how well it meets those standards, its psychometric properties, and how the assessments are administered. This section also provides a summary of the findings of a detailed analysis of the psychometric properties of the Prueba Inicia included as a Technical Appendix, which forms Part II of the report.
- The third section compares the Prueba Inicia to similar exercises in a set of comparator countries. Here the comparison is limited to process and not content, since we would otherwise have had to carry out detailed analysis of the psychometric properties of assessments in other countries. We explore broadly the processes by which teacher licensing or teacher trainees exit exams are conducted in a set of countries¹².
- The final section of Part I, focuses on policy options available to the Government of Chile as it moves to put in place effective measures to screen, recruit and deploy the most effective teachers in classrooms across the country. Overall conclusions are presented, drawing as well on the findings presented in Part II of this report.

Part II of the report provides a detailed technical review of the psychometric properties of the assessment.

¹² Although assessments used for licensing (or certification) are different than assessments used for purposes of selection, at times in this report we use these interchangeably. The overarching objective is teacher selection, and the nuances between assessments for selection and assessments for licensing are at times lost on both the lay reader, and the policy makers, especially when the policies include linking performance on a licensing assessment such as the Prueba, to initial teacher salary scales upon being hired.

IDENTIFYING GOOD TEACHERS

17. **Development policies have been anchored on the importance of investments in human development, particularly through investments in expanding schooling opportunities.** In recent years, this belief that investments in human capital will support the country's growth objectives has come into question as the links between growth and human capital attainment¹³ have not been easily understood. Hanushek and Kimko (2000), Hanushek and Woessmann (2008), Hanushek and Woessmann (2012) illustrate that by employing more direct measures of schooling quality, as opposed to only looking at schooling quantity measure, helps to improve our understanding of cross country variations in long run economic growth. The reasons are quite intuitive. Early models, employing average years of schooling, implicitly assumed that a year spent in a Nigerian school was equal to a year spent in a Singaporean school at the same grade level. Furthermore, it also assumes that everything a student learns is captured entirely by this single measure of attainment, the number of years of schooling, and thus ignoring both the distribution in learning outcomes observed even within a classroom let alone across communities and countries, and all the learning that takes place outside of classrooms.

18. **These findings have spurred policy makers to focus on improving learning outcomes.** The Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) are international efforts to obtain better measures of learning using a set of standardized measures. This global effort to track learning across countries, and to try and understand why children in some countries demonstrate sustained superior results compared to others, stems in part from the increased focus on learning outcomes.

19. **The PISA has captured world-wide attention.** Led by the Organization for Economic Cooperation and Development (OECD), the PISA is aimed at measuring the scholastic performance of 15-year-olds in mathematics, science, and reading. It was first conducted in 2000 and has been repeated every three years since. While cross country comparisons are not the primary objective of the PISA, in the court of public opinion, the PISA has become a direct measure of comparing scholastic achievements across one countries, with tremendous attention given to the best and worst performers on this assessment. The results compel participating countries to introspect - especially those finding themselves at the lower end of the performance curve. Although the OECD makes it clear that PISA results should not be the basis for wholesale changes in educational policies, the global rankings, the associated media fanfare, and public outcry associated with relatively poor performances, often leave policymakers wondering what changes are needed in their respective countries to achieve better student learning outcomes. Teachers and teacher quality always features high on the list of issues to address.

Accountability Measures

20. **Teachers are central to any discussion on education policy irrespective of the country context in which the discussion takes place.** Accountability has increasingly become a buzz word in any discussion on education policy, and is further heightened by the perception that increased expenditures for education have not been met with corresponding improvements education quality. With education quality being measured as improvements in student achievements. A main reasons for this disconnect between increased financing and improvements in learning outcomes is because till recently

¹³ Improved student learning outcomes and quality of education also have spillover effects along many other dimensions including economic competitiveness, productivity, civic participation, or conversely on crime, violence, and other social malaise.

policymakers across the globe were focused on ensuring that children were in school, and not necessarily focused on whether or not they were learning while in school¹⁴.

21. **However, accountability is defined and presented in many ways to cater to the different perceptions and aspirations of the numerous stakeholders involved in the education process.** Students, parents, teachers, school principals, bureaucrats and politicians all view accountability in different ways. Educationists have raised concerns against this seemingly inevitable path that many countries have taken (Ravitch 2013). However, the demand for accountability measures seems to continue its march forward and typically resulting in countries focusing on new measures, such as: (i) extensive student assessment and testing, (ii) using measures of teacher's value added, (iii) ensuring the system's ability to identify high quality inputs – most importantly, teachers, (iv) strengthening access to information and parental participation, and finally (v) linking the flow of funds to all of the above.

Determinants of Teacher Effectiveness

22. **Middle and high income countries, such as Chile, too have begun to adopt such measures of accountability as they transition from focusing on meeting access challenges, to turning their attention towards addressing quality concerns.** Factors which influence student performance have been the subject of research for decades and typically include students' innate abilities, family socio-economic background, parental involvement and support at home, the type and nature of the school and schooling facilities - this includes availability of resources, school and class size, peers, schooling infrastructure, school leadership, and perhaps most importantly, the teacher's role. Gordon, Kane and Staiger (2006) find from studies in Los Angeles between 2000 and 2003 that teachers have a substantial impact on student performance, and that students who had a teacher from the top quarter were likely to be 10 percentile points ahead of their classmates who had a teacher from the bottom quarter of the draw. There are similar findings from other studies as well and increasingly a widespread agreement that teachers make an enormous difference to improving schooling quality (Clotfelter, Ladd, and Vigdor 2007, Clotfelter, Ladd, and Vigdor 2010; McCaffery et al 2004).

23. **The obviousness of the importance of the teacher also arises for three other reasons - budgetary, direct policy control, and contact hours.** From the viewpoint of the policy maker, teachers account for a significant chunk of the allocations made to the education sector through the exchequer every year. In Chile, teacher salaries account for about 80% of the overall education budget in the school sector. Furthermore, of all the levers of change available to policy makers to improve schooling outcomes, only a few are truly malleable. Student ability, characteristics, and family background and circumstances are beyond the reach of governments, while resources, infrastructure, leadership qualities, and teacher policies can be manipulated by policymakers. Finally, the sheer contact time between students and teachers suggests that this is the margin where most effective change can take place. If we assume an average school year of a 180 days and the average number of hours in school per day to be about 6.7 hours, a child spends about 15000 hours in the presence of a teacher between the time she enters Grade 1 and exits Grade 12 (Hattie 2003). It is hard to imagine any serious policy effort that aims to

¹⁴ Global agreements such as Education for All and the Millennium Development Goals have had tremendous impacts on educational outcomes over the last few decades. However, these goals have largely focused on getting children into school and not on learning targets. This is despite the fact that both the Jomtien Declaration of 1990, and the World Education Forums' Framework for Action emphasize the importance of quality education, to ensure that children not only have equal opportunities but also equality in outcomes. The Framework for Action explicitly states that improved quality should lead to recognized and measurable learning outcomes, especially in literacy, numeracy and essential life skills.

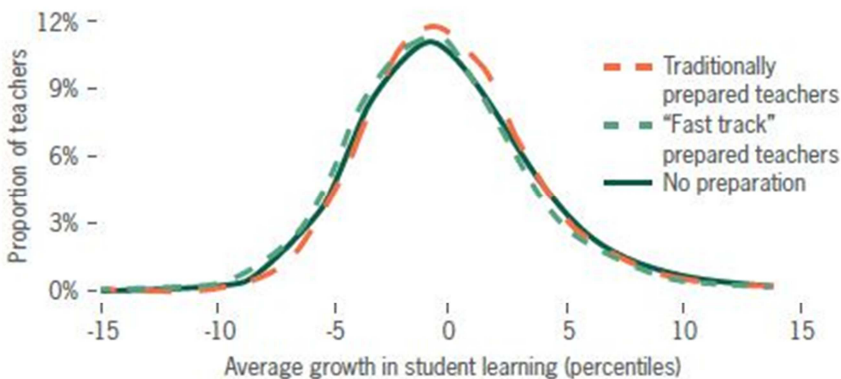
improve schooling quality which does not address or focus on the critical role played by teachers through their daily interactions with students in the classroom.

24. **Two broad streams of thinking have emerged in this area.** One group believes that teacher preparation programs have little or no bearing on teacher effectiveness. Therefore, it might be best to focus less on ex-ante credentials or teacher preparation, and instead focus on lowering barriers to entry and helping teachers get better once they have been brought into the system. This approach would suggest that by manipulating teacher professional development, teacher evaluations, and by delaying (if not completely doing away with) tenure provisions we can ensure that the weakest teachers are weeded out of the system, while the strongest are retained and supported. The second school of thought believes that *all roads do lead to teacher preparation programs*¹⁵ and that while many teacher preparation programs are seen to be performing poorly in many countries, the solution is to strengthen their performance so that the overall standards of teaching in the country can be improved. The following few paragraphs provide some support for both points of view¹⁶.

25. **Gordon, Kane and Staiger (2006) suggest that given that if it is difficult to identify ex-ante those characteristics that make a teacher great, it might be best to rethink the way we pose restrictions on who can join the teaching force and who should not.** The typical system of teacher credentialing involves a deep emphasis on coursework, and proof of teacher content and pedagogical knowledge as seen through test scores. On top of this credentialing process, aspiring teachers typically must also have a bachelor's degree, be licensed or certified¹⁷, and demonstrate competence in their core subject area. However, as their study shows, these paper credentials really tell us very little about how effective or not a teacher is likely to be one placed in an actual classroom. The figure below explains this better.

Figure 2

Does Training Matter?



Source: NCTQ Teacher Prep Review (Copied from Gordon, R., Kane, T.J., and Staiger, D.O., "Identifying Effective Teachers Using Performance on the Job" (Hamilton Project Discussion Paper). Washington, DC: Brookings Institution (April 2006).

¹⁵ Teacher Prep Review – A review of the nation's teacher preparation programs (NCTQ, 2013).

¹⁶ The authors have taken a little bit of liberty in how we interpret some of the messaging coming from the research to help provide an understanding of some of the extreme interpretations that people have drawn from reviewing findings on this matter.

¹⁷ In the US, licensing and certification requirements vary from state to state. While there are efforts to try and make these systems more flexible and have for example, a teacher licensed in the state of Virginia being able to move to Michigan and continue as a teacher there, given that requirements vary considerably implies that teachers usually have to go through another process of licensing and certification.

26. **This figure above illustrates that teacher effectiveness has little to do with the teacher preparation program undertaken prior to becoming teachers.** The figure illustrates that whether a teacher went through all the requirements of a teacher preparation program, or they had been fast-tracked¹⁸, or teachers were brought in with no preparation – had little impact on their effectiveness. They also find that teacher certification reveals very little about how effective they will or will not be. In a study comparing the students of certified versus uncertified teachers in Los Angeles, they found no statistical differences in the achievement scores of children under the two different sets of teachers. They did observe that within each group – certified and uncertified – there was considerable variation in teacher quality¹⁹. These results have been repeated in many other studies. Box 1 below provides a summary of similar findings by the popular author and writer Malcolm Gladwell, who also urges us to consider the fact that since it is difficult to identify using ex-ante measures who an effective teacher is likely to be, then we should consider lowering the barriers to entry and allow for the most effective teachers to be identified and developed through the practice of teaching, and not necessarily based on a written tests as is currently taking place in many parts of the world.

Box 2

The Problem of Identifying Good Teachers Using Ex-Ante Measures

Malcolm Gladwell, best selling author and journalist, raised a storm when he wrote an article entitled *“Most Likely To Succeed – How do we hire when we can’t tell who’s right for the job?”* in the December 15, 2008, issue of *The New Yorker*. Gladwell initially uses experiences from the field of American football and refers to *the quarterback problem* or the inability of coaches and scouts to use a set of indicators or measures to confidently predict the likelihood of success of a player as he transitions from college football to the professional leagues. Gladwell concludes that *“there are certain jobs where almost nothing you can learn about candidates before they start predicts how they’ll do once they’re hired. So how do we know whom to chose in cases like that? In recent years, a number of fields have begun to wrestle with this problem, but none with such profound social consequences as the profession of teaching.”* And, when he brought this otherwise interesting selection issue to looking at the hiring of teachers, he truly created a storm. A key conclusion that Gladwell draws is that if you cannot predict the likelihood of success of an individual based on prior information of someone until they are actually in the job, then it might be best to *have lower barriers to entry to the field even though our initial tendency is to try and actually tighten the standards*. However, the underlying thesis is an interesting one, if you are tasked with selecting someone to become a quarterback or a teacher, and all the prior evidence suggests that it is exceptionally hard to predict winners from others, then what do you do? Gladwell also goes on to say in such situations, what has proven to be very important in ensuring the success of those brought in is the learning environment that exists in these teams or institutions, and how well it supports the development of the chosen candidate.

27. **The National Council for Teacher Quality (NCTQ)²⁰ is an organization that believes that the training of teachers is important to ensuring high quality teachers in their classrooms, but do acknowledge that many teacher preparation programs are operating well below where they need to be functioning.** Box 2 below presents the standards developed by the NCTQ and suggested for teacher preparation programs to support the improvement of teacher quality. The NCTQ prepares standards for teacher preparation programs and believes that if these are fully adhered to then teachers trainees exiting preparation programs would have all the requisite skills to take on challenges inside the classrooms. They believe that teacher preparation – in a rigorous, stylized manner – is essential to ensure higher teacher

¹⁸ For example, with programs such as Teach For America.

¹⁹ Or, that within the certified teachers group – there were good teachers and bad teachers and the same within the uncertified group.

²⁰ In the United States

quality. In a recent report on teacher preparation in the United States, the authors find that teacher preparation in the country is in real trouble. The report uses a four star rating to identify the performance of teacher preparation programs for elementary and secondary schooling. The results are worrying.

Box 3

Quality of Teacher Preparation in the United States²¹

- Some of their findings include:
 - Only about 10 percent of the programs across the country meet the Three-Star rating.
 - There are only 4 programs that meet the Four-Star rating.
 - Only 1 program in the entire country scored above Three-Stars for both the elementary and the secondary level programs.
- The selection of students happens from a much larger pool, and hence weaker pool, compared to students selected in some of the high performing countries like Singapore or Finland- where students selected for teacher preparation programs come from the top third of their graduating classes.

28. **In addition to these, a number of other factors have been studied extensively to determine their impacts on teacher effectiveness and through them on student outcomes.** These are summarized below:

Teacher Education

29. Teacher education is often seen as a starting point for improving quality of education with the assumption being that more qualified teachers would result in better student learning outcomes. Although seemingly logical, studies that have tried to evaluate the impacts of teacher education or qualifications on student learning find the results to be far more nuanced than expected. Earlier studies were hampered by data availability, cross-sectional in nature, and unable to match students with teachers. Hanushek (1986) finds little evidence of observable characteristics, such as, qualifications and experience on student learning outcomes. Hanushek and Rivkin (2006) further confirm this weak link between observable educational qualifications of teachers and student learning outcomes. However, there are some studies which find positive and significant impacts of teacher qualifications on student learning, such as, Betts, Zau and Rice (2003), Nye, Konstantopolous, and Hedges (2004), and Guimaraes and Carnoy (2012), others do not including Enhrenberg and Brewer (1994), Ferguson and Ladd (1996), and Buddin and Zamaro (2009). Therefore at best, one could say that the evidence is mixed.

30. The fact that higher teacher qualifications, in particular, those related to academic or university level programs (e.g., Bachelors or Post-Graduate degree) fail to translate into student learning is counter-intuitive. However, when teacher qualifications is defined on the basis of test scores measuring content knowledge, then some studies have found teacher qualifications to impact positively on student learning, and also support other spillover benefits as well. For example, Enhrenberg and Brewer (1995) observed higher gains in student scores when they were taught by teachers who had scored higher on a verbal aptitude test²². Darling-Hammond (1999, 2000a) finds a positive and significant relationship between teachers who have been trained in the subject matter they then teach in schools and student learning

²¹ Greenberg et al (2013).

²² A key objective of the study was to find the relationship between race and gender, and the race and ethnicity of their students, had little to do with how much students learned. In another paper, the same authors also analyze information from the National Educational Longitudinal Study (1988) and find that once again there was little evidence to suggest that ethnicity and gender had anything to do with student learning outcomes.

outcomes. Similarly, Goldhaber and Brewer (2000) also find a positive relationship between teacher preparation in mathematics and student math outcomes, but they fail to find support for similar findings for science. In terms of spillover benefits, Darling-Hammond (2000b) also finds that teacher attrition from the profession is less for those who have obtained stronger content knowledge training. However, once again this line of research still yields inconclusive findings as numerous studies fail to establish concretely impacts of teachers who have received either subject knowledge training or pedagogical preparation.

Box 4

Standards for the NCTQ Teacher Prep Review

Selection

Standard 1: Selection Criteria.

The program screens for academic caliber in selecting teacher candidates.

Standard applies to: Elementary, Secondary and Special Education programs.

Content preparation

Standard 2: Early Reading.

The program trains teacher candidates to teach reading as prescribed by the Common Core State Standards.

Standard applies to: Elementary and Special Education programs.

Standard 3: English Language Learners.

The program prepares elementary teacher candidates to teach reading to English language learners.

Standard applies to: Elementary programs.

Standard 4: Struggling Readers.

The program prepares elementary teacher candidates to teach reading skills to students at risk of reading failure.

Standard applies to: Elementary programs.

Standard 5: Common Core Elementary Mathematics.

The program prepares teacher candidates to successfully teach to the Common Core State Standards for elementary math.

Standard applies to: Elementary and Special Education programs.

Standard 6: Common Core Elementary Content.

The program ensures that teacher candidates have the broad content preparation necessary to successfully teach to the Common Core State Standards.

Standard applies to: Elementary programs.

Standard 7: Common Core Middle School Content.

The program ensures that teacher candidates have the content preparation necessary to successfully teach to the Common Core State Standards.

Standard applies to: Secondary programs.

Standard 8: Common Core High School Content.

The program ensures that teacher candidates have the content preparation necessary to successfully teach to the Common Core State Standards.

Standard applies to: Secondary programs.

Standard 9: Common Core Content for Special Education.

The program ensures that teacher candidates' content preparation aligns with the Common Core State Standards in the grades they are certified to teach.

Standard applies to: Special Education programs.

Professional skills

Standard 10: Classroom Management.

The program trains teacher candidates to successfully manage classrooms.

Standard applies to: Elementary and Secondary programs.

Standard 11: Lesson Planning.

The program trains teacher candidates how to plan lessons.

Standard applies to: Elementary and Secondary programs.

Standard 12: Assessment and Data.

The program trains teacher candidates how to assess learning and use student performance data to inform instruction.

Standard applies to: Elementary and Secondary programs.

Standard 13: Equity.

The program ensures that teacher candidates experience schools that are successful serving students who have been traditionally underserved.

Standard applies to: Elementary, Secondary and Special Education programs.

Standard 14: Student Teaching.

The program ensures that teacher candidates have a strong student teaching experience.

Standard applies to: Elementary, Secondary and Special Education programs.

Standard 15: Secondary Methods.

The program requires teacher candidates to practice instructional techniques specific to their content area.

Standard applies to: Secondary programs.

Standard 16: Instructional Design for Special Education.

The program trains candidates to design instruction for teaching students with special needs.

Standard applies to: Special Education programs.

Outcomes

Standard 17: Outcomes.

The program and institution collect and monitor data on their graduates.

Standard applies to: Elementary, Secondary and Special Education programs.

Standard 18: Evidence of Effectiveness.

The program's graduates have a positive impact on student learning.

Standard applies to: Elementary and Secondary programs.

Experience

31. Ever since Mincer, experience has become a key tool for those involved in the management of human resources. Market for teachers is no different and both through bureaucratic systems and through teacher union policies, teacher experience still drives numerous factors associated with the training, recruitment, seniority and leadership, career development, teacher transfers, and the structure of compensation. However, the belief that teacher experience is a proxy for teacher effectiveness is fraught with risk. While it is clear that experience is important, it would be too simplistic to assume that all experience is necessarily good or improves the effectiveness of teachers in the classrooms. Earlier studies, such as, by Murnane and Phillips (1981) find a positive relationship between experience and effectiveness, though the results are not statistically significant, nor linear. More recent studies find that teachers just out of teacher training programs are likely to be less effective than teachers who have had some experience Kane, Rockoff, Staiger (2006), and Clotfelter, Ladd and Vigdor (2007, 2010). However, the studies find that this experience premium is short lived. Teacher effectiveness, measured in terms of student learning outcomes, flattens out very quickly over the years. So, although teachers with 20 years of experience on the average are likely to be more effective than teachers with no experience, they are not expected to be significantly more effective on the average than a teacher with about five years of experience (Ladd 2007 and Hanushek 2011). Boyd et al. (2007) finds that in an assessment of gains in math scores, the largest gains happens in the first year and this accounts for about half the cumulative gains seen in terms of the effect of experience for children in Grades 4-5.

Content Versus Pedagogical Knowledge

32. Though the evidence on the impact of subject knowledge preparation is mixed, there is some evidence to suggest that programs of pedagogical support do impact more positively on student learning outcomes. Monk (1994) compares subject matter training versus pedagogical subject training and presents evidence supporting the latter in mathematics. However, even though pedagogical support is considered as absolutely essential, the rigorous evidence is still far from conclusive. However, this is an important issue in teacher preparation programs and we will return to this later in the report.

Teacher Certification and Licensing

33. Teacher certification is important as it is not based solely on teacher's content knowledge but based on in depth assessment of performance in classrooms. This includes an assessment of how well they relate and interact with their students, how well they are able to use available technology to teach students, based on longer term measures of performance often based on self-prepared teacher portfolios that include measures of content knowledge and pedagogical knowledge, and other measures - student assessments, professional development programs, etc. Certification is a process that comes at a later point in time and not when the INICIA is implemented. In many countries an initial certification is awarded at the time the trainee completes the necessary requirements of their teacher education program (such as in Finland) or when licensing exams have been completed, such as, in the US. Certification practices vary considerably across countries and are reflective of political economy concerns rather than based deeply on measures of impact on student learning outcomes. In most countries, the initial teacher certificate tends to be valid for life, though this is not true in the US, Australia and a few other countries. Given the federal structure in the US and Australia, states and territories play a huge role and relocating from one state to another needs teacher certificates and licenses to be revalidated. In most other countries, these procedures tend to be far more centralized and therefore do not raise these concerns. The impact of certification on student learning outcomes is inconclusive. The reasons are many and include

measures of performance for licensing and certification, part-time versus full-time certification, initial certification versus more advanced certification that should be reflective of sustained performance over a period of time.

Teacher Professional Development (TPD)

34. Professional development has become an instrumental part of teacher career development. Although TPD is widely recognized as an important way by which teachers can be motivated to learn and grow, there is very little rigorous evidence to suggest that TPD is instrumental in raising student learning outcomes. In most country contexts, TPD programs take place in an ad hoc manner and it is difficult to gauge the effectiveness of such efforts. There is little rigorous evidence on the impact of TPD on student learning outcomes. Angrist and Lavy (2001) and Jacob and Lefgren (2004) find no impact of TPD on student learning outcomes. Brown et al. (1995) find that focused or targeted TPD programs have positive and significant impacts on student learning outcomes. More recently Harris and Sass (2007) identified what they call the "lagged effect of professional development" and that the benefits of TPD may emerge but not immediately after the training has been completed.

Box 5²³

Building a Better Teacher: How Teaching Works (and How to Teach It to Everyone)

Towards the end of July 2014, Elizabeth Green's book entitled "Building a Better Teacher: How Teaching Works (and How to Teach It to Everyone)" was published and released. The book looks at how some of the best teachers in the United States have taken on this task of improving education in US classrooms on themselves and their ideas and visions on teaching. If one had to distill these wonderfully written 474 pages into three main conclusions – these would be that (i) great teachers are *not* born, (ii) all the skills that make a good teacher can be further deconstructed and each of these skills can then be taught to the next teacher candidate, and (iii) it is important to ensure that well qualified teachers are placed before the students, and not teachers who are not fully prepared. Perhaps the most worrying aspect of the Green's book is that she describes her attempts to teach two classes after losing an argument with a friend and teacher, who states that "*you cannot write a book about teaching if you have never really taught a class*". So, in March 2013, halfway through writing this book, Elizabeth Green stood in front of two classrooms and during the course of the day taught high school social studies. Readers will conclude from her own writing that while she has clearly learned many lessons about teaching, her experience with teaching was near disastrous. And, if this is the case with a well educated individual, whose book would probably end up on the New York Times best-seller list, what should one expect from teachers who are even less prepared to be in front of students in classrooms across the world.

28. **In the case of Chile, reviews of education policy have identified the quality of initial teacher training as one of the main causes for the poor quality of education.** Weissbluth (2013) concludes that one of the key issues is that the state has abdicated its role in the area of teacher training and has placed this entirely into the hands of private entities. Furthermore, programs offering initial teacher training have expanded dramatically over the last ten years, and thus making it difficult for authorities to monitor and help improve the quality of initial teacher preparation. While the selection of good teachers is essential and critical to raising the quality of education, as noted in this section, identifying high quality teachers, using ex-ante measures of performance is extremely difficult, and many of the measures that we can observe do not correlate very well to student learning outcomes.

²³ Elizabeth Green (2014).

THE PRUEBA INICIA AND ITS PROPERTIES

29. This second section really forms the main part of the report. The objective here is to answer the fundamental question of whether or not the Prueba INICIA instrument could be used as a teacher exit exam. If yes, what adaptations might be needed to strengthen the instrument for future use. There are two critical aspects that need to be reviewed – validity aspects and psychometric properties of the tests. This section reviews our findings along both of the above. There are broadly five critical areas that need to be covered in any description or evaluation of such a test. These are: (i) Objectives or Purpose of the Test, (ii) How the Assessment is Developed (what standards are used, validity issues, the development process), (iii) Test Administration, (iv) Psychometric Properties and (v) How are these reported? In this section we attempt to present a summary of some of these issues with respect to the Prueba Inicia.

Objectives or Purpose of the Test

30. **The Ministry of Education in Chile, since 2008, has designed and implemented a diagnostic assessment of knowledge and skills for a career in the education sector.** This diagnostic assessment is mapped to the standards developed for the various subject or content standards published by the Ministry. This helps ensure that universities and other higher education institutions, supporting teacher training programs, improve their initial training programs and the quality of their graduates. Table 3 below shows the number of participants and participating institutions between 2008 and 2011 in the Prueba Inicia assessments.

Table 1
Past INICIA Assessments

Year	Students	Graduates from	And of participating institutions	Tests and other comments
2008	1994	39	49	
2009	3224	43	54	Pedagogy in Elementary and ECD
2010	3616	43	56	Content and pedagogy to Childhood Education, Content and Pedagogical test in basic education
2011	3271	49	59	

31. **Chile has a deep interest in ensuring that all teachers have the requisite content knowledge and teaching skills to be able to become highly effective, high quality teachers.** The Prueba Inicia although at present is a voluntary assessment, there is a desire to mandate this for all teachers on the part of the Government in an effort to ensure that all teachers have the necessary skills and knowledge to be effective teachers. The Prueba Inicia is expected to assess the content knowledge and teaching or pedagogical skills of teacher trainee candidates as they exit their training or teacher preparation programs. Clearly a single dimensional measure of performance on such an assessment is unlikely to help any government determine the future effectiveness of teachers and thus it is anticipated that other measures will be adopted over time to ensure a comprehensive approach.

32. **A battery of tests, collectively constitute the Prueba Inicia covering a wide range of subjects.** As we shall see in the section outlining how these tests are developed, it is believed that each subject test in the Prueba Inicia tests the knowledge and skills of teacher trainees in a manner that reflects what a broad swathe of teachers and professors and other practitioners believe are important content and skills areas for each subject. Since there are now well defined standards developed for each subject or domain, the Prueba Inicia is mapped to these to ensure that the tests are meaningful for teachers as they exit their

teacher preparation programs and that the assessments will eventually support high quality teaching in classrooms since these standards are also expected to be met through instructional practice. The content matter for each subject area of the Prueba Inicia is defined, developed and validated by professionals with domain expertise.

The Development Process

33. **The test development process is a critical aspect of the overall exercise.** In this stage a number of key issues need to be taken into account – (i) test development standards and the manner in which this is carried out, (ii) reviewing and address key concerns regarding validity, (iii) the process of test and item development, and (iv) the piloting and review of tests. We briefly review each of these issues below.

34. **The Prueba Inicia adheres to the guidelines enshrined in the *Standards for Educational and Psychological Testing* by the American Education Research Association (AERA).** The Prueba Inicia’s test development process is anchored in the body of standards noted above. These standards include *inter alia*:

- Clearly defined purpose of the test and claims that can be made about test takers
- Job analysis and content validation/mapping surveys based on content and domains to be tested
- Development of tables of test specifications based on purpose and content
- Develop test items, how many, how are they weighted, etc., based on validation and scope and which measure the behaviors intended
- Formulate complete tests after piloting and reviewing test items
- Ensure fairness or bias concerns
- Ensure that developed assessments do not have problems of overlap or cueing,

35. **The development of the Prueba Inicia is initiated by identifying partner agencies to help in the development of the battery of tests that form the Prueba Inicia.** Given the large battery of tests involved, the overall tasks are typically sub-packaged and divided into several groups and test development activities are contracted out to various parties. The INICIA instruments consist of three sets: (i) tests of Pedagogical Knowledge (PCP), (ii) tests of Knowledge Discipline (PCD), and (iii) a test of Written Communication (PCE)²⁴. These are applied by the graduates at the Preschool, Basic Education level and Secondary Education level. The test instruments cover the following disciplinary areas - Early Childhood Education, Primary Education and Secondary Education in Language and Communication, Mathematics, History, Geography and Social Sciences, Biology, Physics and Chemistry.

36. **Although the development of these tests were contracted out, the overall work was carried out under the technical supervision of the *Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CPEIP)*.** The process is fairly cumbersome and lengthy, but has been well established over the last six years. Initially, universities are asked to propose teacher training standards and this is assessed through a standard pencil and paper type assessment. An iterative process is then adopted with sets of experts to determine whether these standards map well onto the curricular framework and whether these standards are also met by instructional practice.

37. **The proposed standards are then opened up for further discussion and specification test tables are shared and validated through a national consultative process.** Representatives of institutions involved in teacher training at the ECD, Primary and Secondary school levels participate in

24

the process. Public validation of these proposed standards is important in that it provides an opportunity for HEIs to help define the background, key elements, features, and themes needed to help develop the Prueba, and allows these representatives an opportunity to participate on range of issues associated with the assessment.

38. **All the tests are prepared with the same approach.** For example, participants are organized into disciplinary groups – mathematics, physics, etc. For each disciplinary group, discussions are held about the teaching standards and the discipline in question. This is followed by a workshop to analyze the proposal/specification table based on a set of norms, and then each item is given a weight. Finally, this analysis is presented in a plenary session where all are allowed to comment on the set of instruments developed.

39. **Validity is a very important aspect of test development.** It is expected that every time a test is developed there is an intention behind this development. In the case of the Prueba Inicia – the test acts as a filtering device to ensure that students graduating from teacher training programs have the requisite content and skills knowledge to become effective teachers. Clearly other tests may different objectives. For example, the PSU undertaken by Chilean students prior to entry to university is not aimed to set a lower threshold but is intended to identify highly qualified students for intake to university programs. However, irrespective of the objectives of a test, it is critical to ensure that every test *measures what it wants to measure*. Validity is a measure of how well the evidence obtained from these tests supports how the test is to be utilized. This is critical aspect of test development. Though the Prueba Inicia does not call itself a licensing exam, the idea is essentially the same – to ensure that every teacher trainee or test taker –has the requisite content and skills knowledge to perform his or her task as a teacher in an effective manner.

40. **How are validity concerns ensured in test design and development?** To ensure that tests meet these requirements, it is important that the tests fully reflect the content and skills that are deemed to be essential for the particular domain area to be practiced. That is, it is important to ensure that what is being asked of the test taker to demonstrate – knowledge and skills in teaching mathematics for example – must be shown to be important *knowledge and skills to function as an effective mathematics teacher*. For example, a test that assesses a candidate’s skills in higher level areas of math, such as, calculus, trigonometry, vector algebra, etc., would not be an appropriate test for an entry level math teacher for Grade 3. The content to be assessed or tested should be based on the importance of the same in practice of the occupation or profession. Furthermore, it is important to recognize that in the case of the Prueba Inicia, we are really looking to assess *entry-level* skills as in most licensing exams, and thus, this would form only a part of what one might expect from a certification test or a test for example for a master teacher. As noted in the above paragraphs, the mapping content and the test is developed using the expert opinion of teachers in the content area, other practitioners, and key stakeholders using an approach referred to as job analysis.

41. **The table of specification is aimed at developing a comparison and organizing the number of questions mapped to each level of Bloom’s taxonomy.** For example, a math paper may include 20 multiple choice questions, 10 questions on concepts, and 5 questions on drawing graphs. The questions are weighted based on assessments of the degree of difficulty. Once these tables are validated, the teams representing the various universities then develop a wide range of items covering these specification tables. After the CPEIP has approved these items, they are then piloted with a set of teacher trainees across early childhood, primary and secondary schooling. Psychometric analyses is then used to identify and select the best quality items for each axis and then finally, the items are used to join to equivalent test forms which are again finally approved by CPEIP. The key psychometric properties used to narrow down the final list of items included reviewing item discrimination, item difficulty, and non-responses. In addition to this, the two forms are linked by anchor questions and about 15-20 anchor questions are

recommended. Most of the test battery of the Prueba Inicia, except for the test of Written Communication, comprises of multiple choice questions. Each item has four possible options as answers and with a single correct answer with the examinee receiving one point for a correct answer and 0 points for an incorrect answer. Partial scores are not given as the items do not have partially correct alternatives, nor is any correction applied to random responses based off of wrong answers.

Table 2
Prueba Inicia 2013

Domain Areas	Number of Items		
	Form A	Form B	Common
Pedagogical Knowledge Early Childhood Education	50	50	21
Pedagogical Knowledge Primary Education	50	50	23
Pedagogical Knowledge Secondary Education	50	50	19
Pedagogical Knowledge Early Childhood Education	60	60	21
Pedagogical Knowledge Primary Education	80*	80*	31
Content Knowledge for Secondary Education - Language	60	60	24
Content Knowledge for Secondary Education - History	60	60	17
Content Knowledge for Secondary Education - Math	60	60	60
Content Knowledge for Secondary Education - Biology	60	60	60
Content Knowledge for Secondary Education - Physics	60	60	60
Content Knowledge for Secondary Education - Chemistry	60	60	60

Standard Setting

42. **A key step in any assessment of this nature is to determine cut-off scores or score beyond which a candidate is noted to have met the minimum threshold requirements.** This is achieved by conducting studies aimed at setting standards or a cut-off score. Treating the Prueba Inicia as a licensure or credentialing test, would mean that the cut off score is the minimum score that a test taker would have to achieve to be considered as having passed the test and be awarded a license to teach. Cut-off scores have to be able to distinguish between poor or sufficient performance of candidates. Standard setting can be done in several ways but the Prueba Inicia makes use of a normative criterion. That is, it establishes a point in the score distribution and identify that all points above that cut-off point to be acceptable and all below that point to have performed poorly²⁵. In addition to helping establish the minimal level of performance, standard setting also help reaffirm validity of the content as discussed earlier.

43. **Given the nature of the Prueba Inicia – that is, the use of Multiple Choice Questions, the well-established Angoff Method is used for Standard Settings.** The committees of experts, teachers and other practitioners established (as noted earlier) are required to review each item of the test and make a judgement call on what proportion of *expected test takers* would answer the question correctly. At the end of this exercise, the stated proportion for every expert is averaged across items, these judgements are then summed up and averages obtained. This average constitutes the passing score. While the approach is simple, it has several inherent disadvantages. For the Prueba Inicia, an external consultant was engaged to help identify key guiding principles on the basis of which the standards would be established. It is based on this approach that to be classified as having done acceptable on the assessment is required to get a score of 60%. These principles include:

²⁵ Standard setting approaches vary depending on the nature of the assessment and given that the Prueba Inicia is largely an assessments involving Multiple Choice Questions – a normative criterion approach is suitable. This would not be the case for assessment using constructed response items.

- A method based on collective views of experts and to ensure consistency with earlier rounds
- All tests would employ the same approach and in this manner ensure that procedures are standardized and communicated with users
- It should be a method that is appropriate for an exclusive use of multiple choice questions
- Globally accepted practice
- It should be consistent with the statistical model used for psychometric analysis of tests and items

44. **The results for the Prueba Inicia have tended to be very poor.** In 2012, the Prueba Inicia results were termed appalling with over half the candidates who appeared for the examination performing poorly. This was true at all levels – Kindergarten, Primary and Secondary – in which the examination revealed that about 60 percent did not have sufficient understanding in their own subject areas²⁶. The tables below illustrate the poor performance of those who participated in the Inicia in 2012²⁷.

Figure 3
General Results – Prueba Inicia 2012

	Ed. Parvularia		Ed. Básica	Ed. Media
	Niveles	Niveles de desempeño	Niveles de desempeño	Niveles de desempeño
Conocimientos Pedagógicos	Insuficiente	62%	34%	35%
	Aceptable	28%	55%	55%
	Sobresaliente	10%	11%	10%
Conocimientos Disciplinarios	Insuficiente	60%	56%	39% a 76%
	Aceptable	30%	34%	14% a 51%
	Sobresaliente	11%	10%	4% a 11%
Habilidades de comunicación escrita	No Logra nivel adecuado	51%	41%	36%
	Logra un nivel adecuado	49%	59%	64%

²⁶ “Teaching graduates fail national competency test” in the University World News , 7 September 2013 Issue No.286

²⁷ The benchmark figure used by agencies in the US to consider as acceptable teacher training programs is high at 80 percent.

Figure 4
Content Knowledge - Secondary – Prueba Inicia 2012

	Lenguaje y Comunicación	Historia, Geografía, Ciencia Sociales	Matemáticas	Biología	Física	Química
Insuficiente	49%	39%	55%	69%	76%	76%
Aceptable	40%	51%	39%	24%	14%	20%
Sobresaliente	11%	11%	6%	7%	10%	4%

Summary Review of a Sample Test (Example: Mathematics)

45. **In this section we undertake as an example a detailed review of one of the tests used in the Prueba Inicia.** For the purpose of this expositional exercise we use the test entitled *Prueba de Conocimientos Disciplinarios Pedagogía en Educación Media en Matemática* from the INICIA 2012.

Ownership

46. This test is produced and owned by the Ministry of Education, Government of Chile. There are two forms of this particular assessment and these are linked using common items. Both forms of the assessments have a total of 60 questions each in a multiple choice format to be administered in a continuous three hours period. The tests are aimed at entry level secondary school mathematics teachers and are scheduled to be taken towards the end of the teacher preparation program.

Purpose of the Test

47. It is considered to be good practice to state the purpose of the test publicly and have this available for all potential test takers. Although the objectives of the Prueba Inicia assessments are widely shared and known amongst test takers, the test themselves did not identify the objectives at the subject level. For example, it would have been good practice if at the top of the actual examination, the specific purpose of the test could be written. For example, “*The purpose of this test is to measure the knowledge and competencies necessary for an entry level teacher of mathematics at the secondary level*”. A statement of this nature on the examination would be considered good practice.

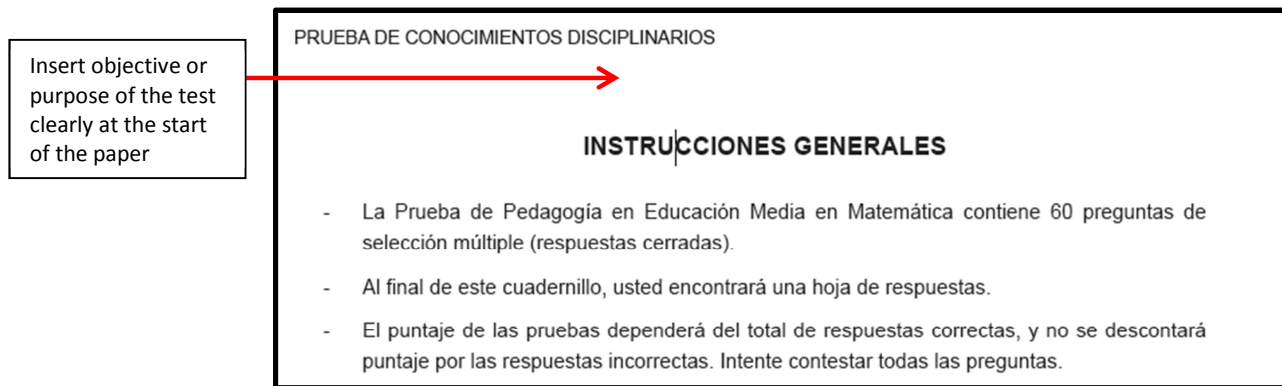
Table of Test Specification

48. As a matter of good practice, prior to developing items, a Table of Test Specifications should be developed. This table helps link the curriculum to be covered and the assessment to ensure consistency. In the Prueba, expert teams reviewed the specifications to ensure alignment between the assessment and the curriculum. When completed, the Table of Test Specifications shows the alignment between the curriculum and the content of the assessment. Of course, it is important that:

- a. the curriculum/standards are clearly mapped to the standards and expected outcomes in each area of content;

- b. weight for each item – as shown through the coverage – maps onto the curriculum and reflects the portion of the curriculum devoted to that topic²⁸.

Figure 5
Good Practice on Test Book Design



49. The areas covered under the assessment are clearly presented in the background document on standards provided along with the Prueba. This is used to develop the Table of Test Specification and the Knowledge/Skills/Abilities matrix. For example, in mathematics, the test consists of five content areas – (i) Number Systems and Algebra (41.7%, 25), (ii) Calculus (5%, 3), (iii) Algebraic Structures (5%, 3), (iv) Geometry (30%, 18), (v) Probability and Statistics (18.3%, 11). The numbers within the brackets show the proportion of each of these in the structure of the assessment, and the second number shows the number of questions under each section. While these detailed breakdowns seem reasonable and appropriate, a content specialist will be needed required to provide a greater understanding of the breakdown²⁹. The table of specifications and the Knowledge/Skills/Abilities matrix were derived in a manner that is consistent with good practice guidelines. The approach involved the use of experts, nominations of key topics and areas to cover by peers, and the use of external reviewers to further validate content and scope. Once the Table of Test Specifications and the KSA matrix were completed, the assessments were piloted to ensure the quality of each item. This is a crucial aspect of any assessment to ensure that overall objectives of the assessment are met and that items were consistent. This approach was used for each of the assessments under the Prueba Inicia and is well documented.

Item Development and Tasks

50. In many cases, development of items can be done using items from question or item banks. Items are chosen carefully to ensure that they meet the purpose and objectives of the assessment, and that the Table of Test Specifications has been used as guide and to ensure that the assessment meets the scope, content rigor and complexity of this table. Item development is a complicated and iterative process, and

²⁸ For example, if 15 percent of the course content is devoted to Algebra and 5 percent is devoted to Trigonometry, then the assessment should not disproportionately test on Trigonometry. This alignment is very important.

²⁹ However, since we are looking for an entry level teacher, and given that topics such as Calculus and Algebraic Structures are typically covered towards the end of the schooling cycle and probably by more experienced teachers, it seems like a reasonable distribution of test items across each content space.

involves numerous rounds of consultations. For example, to develop the assessment for Mathematics, a panel of experts and educators was brought together and determined whether the curriculum content was adequately aligned with the assessment content. There is a lot of subjectivity in the process since these experts would have to view whether the items map well into the instructional aspects of the program, and whether the items are developed with sensitivity. Similar to what is expected of teachers during classroom instruction- in terms of gender, race and ethnicity and similar other individual and idiosyncratic factors. Furthermore, it is important to ensure that the items in an assessment demonstrate content knowledge across the temporal dimension of learning - such as both at the start and end of specific units of learning, make use of the knowledge of experts and educators on common learning issues on specific items, student errors and misconceptions, and some measure of the level of complexity and rigor . Once content area specialists have developed items aligned to the curriculum, and determined the level of complexity, these items would have to be reviewed to ensure that there are clear instructions for each item and to ensure that in the case of multiple choice questions or elected response items, the set of distractors is chosen carefully.

51. **In the case of the Prueba Inicia, test development has followed all the prescriptions for good item development.** The documentation on test development does not state whether or not items were also selected from a test bank that may have existed nor does it focus on whether or not items used in formative assessments through teacher education programs were included. The Government of Chile has painstakingly developed and defined standards across all levels of schooling across a range of subjects. Although the documentation does not state categorically whether or not detailed and well defined standards were in place for all assessed subjects prior to the Prueba's initiation in 2008, it is assumed that this was the case. Concern seems to have been taken to ensure that all items developed for the Prueba demonstrate accuracy and clarity - thus defining the task at hand in a clear and concise fashion, ensure that concepts being measured are well known and understood, ensuring that the item is self-contained and can be solved with the information provided, etc. The quality of the distractors in these assessments, which are typically elected response items or MCQs, is important. The team that has designed the Prueba Inicia seems to have been careful in that not only are the distractors designed to help the assessor learn about routine or typical errors made by examinees, but they have also been designed to ensure that they check any misconceptions in the item, and that the alternatives themselves do not lead to the correct answer. The test development methodology, piloting and the characteristics of the items and tests are reported thoroughly. The documentation is professionally done, exhaustive, and helpful for the next round of test constructors. The relevant units of universities were given the work to do. The reported procedures of the test assembly fulfill the criteria of a professionally-done work: the item writers were selected out of experienced professionals, the test assemblers were professionals, the Table of Specifications were prepared adequately, the relevant stakeholders were involved in the processes or at least they were informed of the processes, the item analysis is done by using proper and adequate practices, and the confidentiality was secured during the process.

Piloting

52. If there was one key aspect of the Prueba that should be or could be faulted it is the issue of piloting the assessment before taking it to scale. While the background documentation of the Prueba Inicia suggests that there was an extensive and thorough process of piloting, this is not evident in the number of items of the actual assessment that have been found to be weak in the detailed psychometric assessment in Part II. In any typical and high quality process, the items would be developed and then piloted to a small sample of potential test takers in an effort to detect and deter possible items problems. As noted earlier, item development is an iterative process. The piloting allows item developers to iron out issues that may have been overlooked prior to actually fielding a large scale and more often than not, costly assessments. For the Prueba, the items developed for the study were administered in formats

similar to the actual test to a representative sample of individuals³⁰. However, from the number of poorly performing items in the actual assessment, it seems as if the pilot was not conducted with the rigor that one would expect. One recommendation from this analysis would be that this process of piloting be carried out with rigor and seriousness so as to ensure that the items are well developed and administered³¹. Though the procedures were adequate in many ways, it seems that the selection of the sample for the piloting was most probably not very successful. The piloting sample was compiled by using volunteer students and teachers. It is known on the basis of the evaluation that there are quite many non-discriminative items. It may be possible that the reason for the low accuracy of the tests lies in the less succeeded sampling in the piloting phase. Additionally, no documentation is found of the final testing, and the related procedures. Hence, it is practically impossible to assess the data management and -analysis or scoring procedure of the final phase.

Validation Concerns

53. As noted above *validity* is an important area of study for an analysis of this nature. There are many different types and forms of validity – content validity, internal and external validity, test validity, construct validity, face validity, etc. There are many ways to assess whether a test or a set of tests really measures what it seems to suggest that it is measuring. The aim of the INICÍA is "to monitor the knowledge and skills of new graduates from pre-teacher training institutions". It is quite obvious, that the tests measure the knowledge dimension of the new graduates and it gives only a restricted picture of the skills of the graduates. Such dimensions of a good teacher as the personality of the teacher, pedagogical skills in action, and classroom management are measured in lesser or nonexistent quantity. In the paragraphs below we examine some of these in more detail.

- a. Face Validity: Face Validity is a measure of the representativeness of a research project, and whether it appears to be a good project. From the face validity viewpoint, the tests are interesting, professional looking, and versatile though restricted to Multiple Choice type of questions. The reports describing the procedures of developing the instruments show that the work was done professionally and seriously. To make the tests even more versatile, a couple of productive items would raise the standard. The aim of the *INICÍA* examination is "to monitor the knowledge and skills of new graduates from pre-teacher training institutions". It is quite obvious though, that the tests measure the *knowledge* dimension of the new graduates, but it is not clear that the assessments measure the *skills* dimension of the graduates.
- b. The structures of the tests are well-documented by the test developers, they are based on a relevant theoretical framework (school curricula), and the observed structure correspond with the aimed one. Hence, the structures of the tests seem valid. However, maximizing the validity over the reliability may be one reason why the *reliabilities* of the sub-tests of *INICÍA* are quite *low*. The number of linking items is proper for the stable estimation of the items parameters over the versions. The contents of the tests were

³⁰ In practice, it is urged that the items to be evaluated and the manner in which it is evaluated replicate to the extent possible the actual assessments. This implies that items being evaluated, and the actual administration of the assessment, should be as close as possible to the actual assessments – in terms of content, structure, administrative process, student population (so, as to ensure that there is similarity across student motivation, preparation, item difficulty, etc.)

³¹ There are of course, additional complications. Even if items are statistically well developed, this does not guarantee that the test is a good one since the assessment could be using poorly aligned content, but which perform well statistically.

based on either the national curricula or the Guiding Standards for Educational and Alumni Career in Basic Education, Early Childhood or *Media*. Hence, there is no doubt that the contents of the tests are valid to measure the knowledge base of the beginning teachers.

- c. From the content validity viewpoint, the contents of the tests were based on either the national curricula or the *Estándares Orientadores para Egresados de Carreras de Pedagogía en Educación Básica, Parvularia o Media*. Hence, there is no doubt that the contents of the tests are valid to measure the knowledge base of the beginning teachers. An exhaustive analysis of the contents would need quite may substance experts.
- d. From the viewpoint of ecological validity, the depth of the tests is versatile for testing the cognitive processes of the graduate teacher. The proportions of Knowledge-, Comprehension-, and Higher skills items were fixed to 30%, 40% and 30% respectively. Intuitively, the number of recall-type of items feels quite high in comparison with the international practice; the international student assessment settings as PISA and TIMSS seem to be geared toward application rather than memorizing things. In *INICÍA*, the Application and Higher skills seem to be combined though it seems, however, that these items are geared toward Higher skills even though they are called “skill-related items”.
- e. From the viewpoint of structure validity, the structures of the tests are well-documented by the test developers, they are based on a relevant theoretical framework (school curricula), and the observed structure correspond with the aimed one. Hence, the structures of the tests seem valid. However, by maximizing the validity over the reliability may be one reason why the reliabilities of the sub-tests of *INICÍA* are quite low. The reliabilities for high stake tests are high or sufficient only in the tests of *PCD-Física* ($\alpha = 0.91$) and *PCD-Matemática* ($\alpha = 0.88$). The number of linking items is proper for the stable estimation of the items parameters over the versions.

54. In summary, the *INICÍA* examination seems to be professionally developed, adhering to well established and used set of standards, they are versatile and motivating though restricted in their measure with an emphasis on the knowledge aspect of the graduating teacher. The *INICÍA* examination is very limited from some other relevant aspects of the “good teaching”, such as the classroom management, pedagogical skills, or personal traits of the graduates.

Figure 6
Mathematics Standards Table for Secondary Schools

Topics	Standards	N° Items. Percentage
Number Systems and Algebra	1 is able to drive the learning of number systems N, Z, Q, R and C. 2 is capable of conducting operations learning of elementary algebra and its applications for solving equations and inequalities. 3 is able to drive the learning of the concept of function, their properties and performances. 4 shows disciplinary competence in linear algebra and is able to drive learning applications in school mathematics.	25 Questions 41,7%
Calculus	5 is able to drive the learning of real numbers, sequences, and series summations. 6 shows disciplinary competence in differential calculus and applications. 7 shows disciplinary competence in integral calculus and applications..	3 Questions 5%
Algebraic Structures	8 is able to drive learning divisibility of integers and polynomials and demonstrates disciplinary competence in its generalization to the ring structure. 9 Demonstrates competence in disciplinary theory of groups and bodies. 10 shows disciplinary competence in basic concepts and constructs of mathematics..	3 Questions 5%
Geometry	11 is able to drive the learning of basic concepts of geometry. 12 is able to drive learning and homotecias isometric transformations of figures in the plane. 13 is able to drive the students' learning on issues related to measurement of geometric objects and attributes using trigonometry. 14 is able to drive the learning plane analytic geometry. 15 is able to drive the learning of geometry using vectors and space coordinates. 16 Includes foundational aspects of Euclidean geometry and some basic models of non-Euclidean geometries.	18 Questions 30%
Probability and Statistics	17 is able to motivate the collection and study of data and conducting learning the basic tools of their representation and analysis. 18 is able to drive the learning of discrete probabilities. 19 Ready to drive learning of discrete random variables. 20 Ready to drive learning normal distribution and limit theorems. 21 Ready to drive learning of statistical inference..	11 Questions 18,3%

Documentation

55. For any assessment of this nature, it is important to ensure that the test takers have all the information that is needed for an assessment. The documentation available for the Prueba Inicia is weak in several ways – (i) the entire process of test development even if well documented is not easily available for example, the technical details of the test for each subject, the results of the pilots, etc.,(iii) information for all those ready to take the assessments – for test information, test preparation and the actual test itself. Although as stated earlier, the Prueba Inicia exams seem to have been developed professionally, when placed in comparison to assessments such as the Praxis series assessments look more professionally developed. For example, the Praxis series documentation is fully available on line with clear documentation on the tests themselves, their coverage, there are test preparation material available online for free, and there are study guides that help you prepare for the test. In addition, there are also available online assessments for practice that allows the candidate to the extent possible mimic the actual settings of such exams. Familiarity with the test methodology is an important aspect in any attempt at a test. Even for content knowledge specialists, with a high degree of content knowledge, changing test formats and making testing styles very different will impact on their scores. The reason why access to such documentation and support material is important – especially if this is made a high-stakes mandatory assessment – is because the aim and purpose of the assessment is to obtain an understanding of the

content and skill knowledge of a trainee in a particular subject. If performance is affected because the student is not familiar with certain aspects of the test that would add unnecessary noise to the assessment. For example, there are significant differences between *paper based assessments* and *computerized assessments*, or whether or not the assessment is a mix of multiple choice questions or constructed response questions.

56. Therefore, the documentation surrounding the Prueba Inicia assessments could be strengthened – in terms of content and presentation. A simple example is that the current version of the assessment does not: inform the examinee:

- a. Inform the examinee about how long they have to complete the assessment
- b. Inform the examinee about the manner in which the marking will be done, for example, is there negative marking or not? H
- c. Have the usual guidance notes on “not getting stuck with a difficult question” and that you should answer as many as you can.
- d. There is a page that provides indicative notes to the examinees for mathematics – under the title *INSTRUCCIONES ESPECÍFICAS*. However, following this page the actual assessment begins and there is nothing that illustrates this to the examinee. The figure on the following page illustrates this and compares the Prueba Inicia paper to an actual Praxis series assessment both in Mathematics.

57. Access to test documentation as we have noted above is an important feature of any assessment of this nature. While these may appear to be small issues, efforts should be made to ensure that the Prueba Inicia follows and adopts best practice norms on issues like these.

Figure 7
Comparing The Praxis and Prueba Mathematics Examination

From Prueba Math Assessment	From Praxis Math Assessment								
<p>SÍMBOLOS MATEMÁTICOS</p> <table style="width: 100%; border: none;"> <tr> <td style="width: 50%; border: none;">$[a, b]$ intervalo cerrado</td> <td style="width: 50%; border: none;">$]n, b[$ intervalo abierto</td> </tr> <tr> <td style="border: none;">$[a, b[$ intervalo cerrado por la izquierda y abierto por la derecha</td> <td style="border: none;">$]n, b]$ intervalo abierto por la izquierda y cerrado por la derecha</td> </tr> <tr> <td style="border: none;">\cong es congruente con</td> <td style="border: none;">\sim es semejante con</td> </tr> </table> <p style="text-align: center;">3 CDEM-M2</p>	$[a, b]$ intervalo cerrado	$]n, b[$ intervalo abierto	$[a, b[$ intervalo cerrado por la izquierda y abierto por la derecha	$]n, b]$ intervalo abierto por la izquierda y cerrado por la derecha	\cong es congruente con	\sim es semejante con	<p style="text-align: center;"><i>Integration by Parts</i></p> $\int u dv = uv - \int v du$ <p style="text-align: center;">-8-</p> <p style="text-align: center; font-size: small;">This ebook was issued to _____ order #9616612238. Unlawful distribution of this ebook is prohibited.</p> <hr/> <p style="text-align: center;">MATHEMATICS: CONTENT KNOWLEDGE</p> <p style="text-align: center;">Time—120 minutes 50 Questions</p> <p style="text-align: center; font-size: x-small;">Directions: Each of the questions or incomplete statements below is followed by four suggested answers or completions. Select the one that is best in each case and then fill in the corresponding lettered space on the answer sheet with a heavy, dark mark so that you cannot see the letter.</p> <table style="width: 100%; border: none;"> <tr> <td style="width: 50%; border: none; vertical-align: top;"> <div style="border: 1px solid black; padding: 2px; text-align: center; margin-bottom: 5px;">1.8984375 E -20</div> <p>1. Shown above is a number displayed in scientific notation on a calculator. What is the 20th digit to the right of the decimal point when the number is expressed in decimal notation?</p> <p>(A) 0 (B) 1 (C) 5 (D) 8</p> </td> <td style="width: 50%; border: none; vertical-align: top;"> <p>2. Juanita is 4 feet tall. Which of the following could be used to calculate her height in centimeters? (1 inch = 2.54 centimeters)</p> <p>(A) $4 \text{ ft} \times \frac{1 \text{ ft}}{12 \text{ in}} \times \frac{1 \text{ in}}{2.54 \text{ cm}}$</p> <p>(B) $4 \text{ ft} \times \frac{1 \text{ ft}}{12 \text{ in}} \times \frac{2.54 \text{ cm}}{1 \text{ in}}$</p> <p>(C) $4 \text{ ft} \times \frac{12 \text{ in}}{1 \text{ ft}} \times \frac{1 \text{ in}}{2.54 \text{ cm}}$</p> <p style="text-align: right; font-size: x-small;">17 in 7 54 cm</p> </td> </tr> </table>	<div style="border: 1px solid black; padding: 2px; text-align: center; margin-bottom: 5px;">1.8984375 E -20</div> <p>1. Shown above is a number displayed in scientific notation on a calculator. What is the 20th digit to the right of the decimal point when the number is expressed in decimal notation?</p> <p>(A) 0 (B) 1 (C) 5 (D) 8</p>	<p>2. Juanita is 4 feet tall. Which of the following could be used to calculate her height in centimeters? (1 inch = 2.54 centimeters)</p> <p>(A) $4 \text{ ft} \times \frac{1 \text{ ft}}{12 \text{ in}} \times \frac{1 \text{ in}}{2.54 \text{ cm}}$</p> <p>(B) $4 \text{ ft} \times \frac{1 \text{ ft}}{12 \text{ in}} \times \frac{2.54 \text{ cm}}{1 \text{ in}}$</p> <p>(C) $4 \text{ ft} \times \frac{12 \text{ in}}{1 \text{ ft}} \times \frac{1 \text{ in}}{2.54 \text{ cm}}$</p> <p style="text-align: right; font-size: x-small;">17 in 7 54 cm</p>
$[a, b]$ intervalo cerrado	$]n, b[$ intervalo abierto								
$[a, b[$ intervalo cerrado por la izquierda y abierto por la derecha	$]n, b]$ intervalo abierto por la izquierda y cerrado por la derecha								
\cong es congruente con	\sim es semejante con								
<div style="border: 1px solid black; padding: 2px; text-align: center; margin-bottom: 5px;">1.8984375 E -20</div> <p>1. Shown above is a number displayed in scientific notation on a calculator. What is the 20th digit to the right of the decimal point when the number is expressed in decimal notation?</p> <p>(A) 0 (B) 1 (C) 5 (D) 8</p>	<p>2. Juanita is 4 feet tall. Which of the following could be used to calculate her height in centimeters? (1 inch = 2.54 centimeters)</p> <p>(A) $4 \text{ ft} \times \frac{1 \text{ ft}}{12 \text{ in}} \times \frac{1 \text{ in}}{2.54 \text{ cm}}$</p> <p>(B) $4 \text{ ft} \times \frac{1 \text{ ft}}{12 \text{ in}} \times \frac{2.54 \text{ cm}}{1 \text{ in}}$</p> <p>(C) $4 \text{ ft} \times \frac{12 \text{ in}}{1 \text{ ft}} \times \frac{1 \text{ in}}{2.54 \text{ cm}}$</p> <p style="text-align: right; font-size: x-small;">17 in 7 54 cm</p>								
PRUEBA DE CONOCIMIENTOS DISCIPLINARIOS									

Psychometric Properties

58. The attached technical section takes a deep look at the psychometric properties of the Prueba Inicia. This uses a variety of psychometric and statistical techniques and adheres to established standards for such assessments. The data for the analysis comes from the Ministry of Education, Government of Chile, are the actual responses from individual test takers to each item in the administered test forms. The data sets are analysed in five ways: (i) classical test theory and item analysis, (ii) similar analysis is also used to determine items which perform poorly, (iii) a DIF analysis is conducted to ensure that the Prueba meets typical standards for fairness with high DIF scores suggesting that items may need to be reviewed, (iv) using IRT modelling to calibrate items across the tests in to a common scale and acquire item difficulty levels that can be compared, and (v) IRT modelling to equate test scores across different tests and by doing so determine whether the original scores are comparable across tests. This is very important to ensure that when looking across tests, whether the cut off boundaries of Outstanding, Sufficient and Insufficient continue to be comparable across tests of maths, language, etc. The results from this analysis are summarized below.

Item Analysis

59. **Given that the Prueba Inicia is intended to be a high stakes assessment - the reliabilities of the sub-tests of the Prueba Inicia might be considered to be low in many cases.** The reliability of the scores reflects strictly the accuracy and discrimination power of the test; *the lower the reliability the less accurately the total score reflects the true ability of the test-takers.* The reliabilities for each subtest is denoted by α are shown here, such as, $\alpha = 0.64$ (PCE-INICÍA), $\alpha = 0.66$ (PCP-Basica), $\alpha = 0.68$ (PCP-Parvularia), and $\alpha = 0.69$ (PCD-Parvularia) are very low and $\alpha = 0.71$ (PCD-Lenguaje), $\alpha = 0.72$ (PCP-Media), $\alpha = 0.74$ (PCD-Historia), and $\alpha = 0.77$ (PCD-Biología). In many cases, the standard error of measurement is more than ± 3 points which leads to a situation in some tests that the “insufficient” and “outstanding” test-taker can be reversed. For what is eventually intended to be a summative, high stakes assessment as the Prueba Inicia, items with such low reliability will pose problems.

60. **From the earlier sections we noted that the development of the Prueba Inicia seems to have met all the established standards for test development, and yet a surprisingly large number of items are found to have low discrimination.** Despite the efforts that have clearly gone into the preparation, development and design of the Prueba Inicia tests, the final INICÍA test set includes a fair number of low-discriminating items. Out of 915 items, there are 19 (2.1%) pathological items with negative item-total correlation and 294 (32.1%) of those which should have been omitted at the final phase because of very low items discrimination ($R_{it} < 0.20$). Given the intended purposes of the test, it would be important to either to omit or rewrite these to raise the standard of the tests or select new items.

61. **The key problems with the items can be characterized as: (i) only on real alternative to select, (ii) multiple possible answers might prove to be correct, (iii) low ability students seem to be able to guess the right answer and (iv) negative item correlation.** There seems to be four kinds of challenges in the flagged items. In many cases, there is *only one alternative to select* – which happens to be the correct one. In these items, even the weakest students know, just recognize, or guess the correct answer too easily and, hence, the low item discrimination. In these cases there are also usually one or more alternatives which are never selected. It may be worthwhile to rewrite the items so that these alternatives are amended, if possible, to more attractive so that the weakest students would select those distractors. Another commonly seen challenge is that there seems to be *several “correct” answers* which attract the best students. The main law is that the best students should select the correct alternative more probable than the weaker ones. In many items of the INICÍA, this does not happen. It may be worth

considering revising (or at least checking) the items so that there really are not those kinds of alternatives which can be (partly) correct ones according to the latest results of the latest journals, for example. Two less common challenges are connected by the fact that the *weakest students seem to guess the correct answers too easily*. In some cases, this evidently leads to the pathological, *negative, item-test correlation*. The latter may be caused also the fact that there seems to be several items where the graphical analysis suggests that the key was not correct. Obviously, these items should be omitted or rewritten.

62. **In any assessment, there is a mix of easy items and difficult items. The Prueba Inicia seems to be more geared towards easy items.** From the IRT modelling viewpoint, the difficulty levels of the items (B parameters in IRT modelling) range from $B = -4.082$ to $B = 3.14$. The distribution of the item difficulties is geared toward easier items rather than difficult items. From the test construction point of view it would be good if the really good test takers had been given an opportunity to show how good they are. Now it seems that each three most difficult item (Bio_A47, Bio_A40, and His_A40) are flagged as pathological ones; the item discrimination is negative and the percentage of correct answers is $p < 0.04$. The reason may be an incorrect key.

63. **As stated above, fairness is an important feature of any assessment. The Differential Item Functioning (DIF) analysis is carried out to ensure that the items meet all the standards for fairness.** By comparing the performance across different subgroups of test takers, say by gender, it is possible to see the responses of male and female and these subgroups compared. High DIF statistics suggests that another look at the item might be warranted. The Mantel-Haenszel statistic (MH) and a graphical evaluation were used to assess the DIF of the tests. The number of cases is, in most datasets, too sparse to perform a proper DIF analysis even for the smallest number of the comparable groups, that is, when comparing two groups. However, the DIF of the items were tested on the basis of the variable *Tipo de evaluado* which has two values: 1=*Egresado de pedagogía* and 2=*Beca Vocación de Profesor o Enseña Chile*. MH gives the result as the Standard Normal distribution fractions. Statistically significant DIF would require values over 1.96. None of 915 items showed this high value. Hence, from the statistical viewpoint, none of the items show DIF. The graphical analysis, however, shows grave discrepancies between the groups.

64. **A concerning aspect of the Prueba Inicia's design seems to be that the individual tests and test versions are not a the same level of difficulty and therefore not all subject or domain areas are being treated equally.** A key concern emerging from the assessment is whether across subject areas, the individual tests are assessed in a similar manner and whether the reporting categories of *Insufficient*, *Sufficient*, and *Outstanding* are fair for all test takers. It is evident that the individual tests and test versions are not at the same difficulty levels. This should have been addressed when constructing the reporting categories. To put this into perspective, the mediocre test-taker with the latent ability of $\theta = 0.00$ would gain in the *PCD-Física* only 31 points while with the same latent ability level, a test-taker in the *PCD-Historia* or in *PCD-Lenguaje* would gain 40 points even though the maximum values of the tests are the same. The latter tests being far easier than the former one. The proper approach would have been to equate the scores before calculating the reporting categories³².

³² The challenge in the reporting categories is that they are based on a set of norm-referenced tests. Hence, there are no absolute criteria as to where to set the boundaries for "insufficient", "sufficient", and "outstanding" test-takers. The relevant question then becomes, who decides where the boundaries are and on what is the basis for making such a decision? In the norm-referenced testing, it may happen that all the candidates are good enough in an absolute sense but the norm always points out some test-takers to be the lowest ones and the others to be the highest ones. Hence, the boundaries for "insufficient", "sufficient", and "outstanding" are not fixed in an absolute sense.

65. **The comparability of the standard deviations urges the equating of the test scores.** As noted in the previous paragraph, equating the test scores before determining the boundary conditions would have been more appropriate than what was used in the Prueba Inicia. It would be better to equate the test scores over the tests and to use the latent ability (Theta) as the indicator for the cut-offs rather than standardizing the scores within the single test. Equating would cause the boundaries to be comparable over the different tests of different difficulty levels. Furthermore, the standard errors of the measurements are high, and the ranges from “insufficiency” to “outstanding” seem too narrow to make a distinction between the test-takers³³. This essentially implies that a test taker at the upper boundary of “Insufficiency”, if taking the test on another day could be labelled as “Outstanding” in the tests of *PCP-Parvularia*, *PCD-Biología*, and perhaps even *PCD-Parvularia*. This is because the labelling system is not coherent across the tests. This approach is not appropriate for the Prueba Inicia, since when rolled out to full scale, this will be a high stakes licensure exam³⁴.

66. **In the tests of the Prueba Inicia, the boundary conditions for failing or insufficiency are also seen to have been set high.** The boundaries for “insufficiency” or “failing” are set relatively high. For example, in *PCE-INICIA* the boundary for failing is set to 50% of the maximum score, in *PCD-Básica* one needs to reach 59% of the total score in order to be “Sufficient”, in *PCD-Biología*, *-Historia*, and *-Parvularia* 60%, in *-Física* 63%, in *-Matemática* and *-Química* 65%, and in *Lenguaje* as high as 68%. Hence, the requirements for being “sufficient” are quite high. Another option, used in the studies of “weak” students, is to use the criterion of 1.5 standard points below the average as the benchmark.

What do teachers have to say about the Prueba Inicia

67. **Meckes et al (2013) note that the accountability regime in Chile has shifted from an era of very low stakes accountability measures to a time of high stakes accountability³⁵.** The authors use two separate years to review how teacher training institutions and the candidates themselves modified behavior in an effort to undertake these assessments. Their paper allows us to answer the following questions: (i) how did your institution or you prepare for the Inicia?, (ii) did anyone put pressure on you to perform well in the assessment, if yes, who was this? The study also looks at the perception of Inicia by students, heads of training institutions.

42. **Pressure to Participate and Do Well in the INICIA:** The study finds that about half the candidates perceived pressure from their institutions to perform well and the remaining half felt pressured by themselves to take the assessment seriously. Students in high performing institutions did not feel as much pressure from authorities, as did those from more poorly performing programs.

43. **Test Preparation:** The study finds again that students belonging to high performing institutions were less likely to receive assistance for test preparation while about 80 percent of the students in lower performing institutions received some form of test preparation assistance. Furthermore, between 2010

³³ Refer to the next paragraph on the lower bounds being set too high.

³⁴ By using the rule of “ ± 1.5 std. units” would not lead to the situation where the true abilities of the “insufficient” and “outstanding” could be the same.

³⁵ Although the assessment began in 2008, they have remained voluntary. Even then participation rates at the level of institutions have been high. In 2013, only 14 percent of total eligible candidates participated in the assessment. Furthermore, for the first couple of years, the results at the institution level were not released to safeguard institutional reputation and to encourage institutions and candidates to participate in the programs.

and 2011, we see an across the board increase in the employment of test preparation exercises between these two years³⁶.

44. Program directors report results that do not fully conform to the findings from student. The more successful programs also claimed to provide assistance to students for taking the test. Furthermore, even though the study finds that both institutions performed relatively similarly, their responses to the Inicia results were vastly different – with one institution internalizing the poor performance and trying to find ways to improve, while the other institution felt that the results were due to the nature of the assessments and that their poor performance was due to external factors.

Three Key Concluding Points

68. **There is a need to take a look at the instruments of the Prueba Inicia from the view point of reliability.** The key technical challenge in the INICÍA lies in its low accuracy. The overall reliabilities could be considered as being low (in most tests, $\alpha < 0.75$) for what is eventually intended to be a universal, high stakes assessments. The tests include too many low-discriminating items and even some pathological items. In some cases, just checking whether the key is correct may help address the problem. However, if this fails to do so, it might be important to omit/rewrite the pathological and poor items and this would raise the standards of the assessment considerably.

69. **While the reporting categories for the Prueba Inicia have been found to be adequate, there is room for criticism of their boundaries.** As noted in the previous section, the test scores do need to equated across test areas, this is particularly important to ensure consistency across the different subject assessments. Furthermore, the boundaries for "Insufficient", "Sufficient", and "Outstanding" needs to be reassessed. The range from "Insufficiency" to "Outstanding" is too narrow in some tests compared with the standard error of measurement. Another systemic of " ± 1.5 standard units" related to equated scores could be considered; this would lead to such boundaries as "exceptionally low" and "exceptionally high". The concept of "Insufficiency" needs to be reviewed carefully; the norm-referenced testing does not provide such indicators that could be used as benchmark for the "Failing" - the labels of "Failing" or "Insufficient" should be used cautiously.

70. **The Inicia test set is a good set of tests of content knowledge of a graduating teacher trainee, however, from the view point of ecological validity, one could ask whether this is sufficient.** From the view point of validity, the INICÍA test set is a good set of tests for measuring and assessing the content knowledge of students graduating from teacher training institutions. The assessments are versatile, the test forms are well developed and the items are made to look interesting, the instructional material on the test forms are very clear, the contents and coverage seems adequate across all subjects. The validity challenge comes from the ecological aspects of validity: does the test really measure the skills needed in the real life teaching? Though content knowledge is clearly of importance it assesses only one dimension of teaching. Teaching methods and skills are clearly important and perhaps are more important in earlier grades than perhaps in higher grades since the composition of the expected levels of student self-study increases in higher grades. Furthermore, again classroom management skills are clearly important and it is important to understand how best to assess this aspect of teacher preparation. The Prueba would benefit tremendously by introducing elements that measure teaching skills and performance more directly.

³⁶ The study finds that in some cases institutions actually threatened their students to participate in the assessment and failure to do so would have been punished either in the form of additional tasks being needed to be completed or in terms of losing fee refunds or having to pay more.

BENCHMARKING THE PRUEBA INICIA

45. **In this section we will review the INICIA assessment and compare it with procedures employed in a select set of countries**³⁷. Benchmarking³⁸ of this nature is employed to understand and incorporate successful lessons from similar exercises in other settings and to understand why some processes, procedures and instruments in existing programs might work while others need to be modified. Given that INICIA-like programs have been implemented in other countries and for longer periods of time, provides an opportunity to identify issues or areas of concern that might arise, and modify or adapt different procedures in INICIA to improve the procedures by which thresholds can be established for candidates exiting teacher preparation programs. It is important to note that in this section, we are benchmarking *processes*, not the psychometric properties of the INICIA, as this would have required us carry out assessments similar to that shown in Part II for several other countries, and this is beyond the scope of the project.

46. **The INICIA is not a unique model and globally there are many similar teacher pre-service assessments.** As noted earlier, the preoccupation with teacher quality is indeed a global issue. It is an issue in those countries who students perform poorly in terms of learning outcomes, and in those countries where their students perform well and score high in terms of student learning. This concern stems in part from the recognition that teacher quality can have an enormous impact on student performance, and has the ability to erase the gap in performance seen across different groups of students or even across countries.

47. **Tightening standards to improve the quality of teachers.** This desire to improve the quality of schooling, particularly at a time when students numbers in all countries have grown dramatically is the main reason why governments have pushed legislation to try and tighten the standards for becoming teachers. As we have noted earlier, there a number of factors that govern teacher quality including: entry criteria, teacher preparation/education programs, selection mechanisms into the profession, teacher professional development opportunities, tenure, compensation and incentive mechanisms, etc. While the factors that come into play in the post-selection/recruitment phase – such as, teacher compensation, teacher professional development, etc. are important³⁹, this report has focused on the phase of teacher preparation. We limit our attention to this preparation phase. Specifically, we look at the following key factors: (i) intake into programs, (ii) exit requirements from teacher preparation programs, (iii) licensing and/or certification, (iv) content versus pedagogical knowledge, (v) performance based assessments, (vi) curriculum – both in teacher preparation programs and more generally the established set of standards across countries, and (vii) quality assurance mechanisms supporting all processes till teacher selection.

³⁷ As part of the benchmarking exercise, we compare specific features of the INICIA with how similar issues are handled and dealt with in other countries. Although in a typical benchmarking exercise, there is a tendency to look at best practices, in this note we refrain from making comparisons with the best practice given that many interventions being experimented with across the world are relatively new, and the jury is still out on the effectiveness of these interventions.

³⁸ Instead of using a fixed set of countries against which to benchmark, we have instead decided to look at specific issues that are relevant for teacher preparation programs and then look at how countries differ. Therefore, while on the issue of teacher intake we may, for example, compare Chile to Singapore and the US, on teacher licensing and certification, we may or may not compare Chile to Singapore and the US, but instead compare the across states of the US and the UK. This way a greater set of issues can be covered without necessarily tying the report to how particular countries undertake these activities.




³⁹ Although we have taken this approach to look at direct factors, indirect factors such as compensation packages, likelihood of tenure, pension programs, etc. all also impact on the decision to try to become or not a teacher. The fact that most of the top students in most countries opt out of teaching at the school level suggests that other factors are clearly at play, but in this review we do not take these issues into account.

Quality Assurance Mechanisms

48. **Mechanisms for quality assurances.** Tatto et al (2012) illustrate that Chilean teachers do very poorly both in terms of subject knowledge and in terms of pedagogical knowledge in teaching mathematics. Figure below shows a comparison of Chilean mechanisms for quality assurance with those of 16 other countries and fall well short of how these countries perform. As stated above, the poor quality of teacher training programs can be attributed to three main factors: (i) rapid expansion of the program and increase in number of people seeking teaching as a career, (ii) a low barrier to entry and the typical student who makes it in and out of a not too selective program for teacher training is most likely to end up catering to the weakest and poorest members of the country.

Figure 8
Mechanisms for Quality Assurance

Country	Entry into Teacher Education			Accreditation of Teacher Education Programs	Entry to the Teaching Profession	Relative Strength of Quality Assurance System
	Control over supply of teacher education students	Promotion of teaching as an attractive career	Selection standards for entry to teacher education			
Botswana						Moderate
Canada						Moderate
Chile						Low
Chinese Taipei						High
Georgia						Low
Germany						Moderate/High
Malaysia						Moderate
Norway						Moderate/Low
Oman (Secondary)						Low
Philippines						Low
Poland						Moderate
Russian Federation						Moderate
Singapore						High
Spain						Moderate/Low
Switzerland						Moderate
Thailand						Low
United States						Moderate

Key:  Strong quality assurance procedures  Moderately strong quality assurance procedures  Limited quality assurance procedures

Source: Tatto et al, (2012)

Teacher Intake

49. **Intake into teacher education programs will have considerable impacts on the quality of teachers eventually produced.** Entry to teacher education programs vary considerably across countries. There are several important issues to consider including: (i) at what level does entry happen, (ii) manner in which entry into teacher preparation programs are controlled, (iii) the population from which most of the students are sourced. Entry into teacher preparation programs happen at many different levels across countries, although in almost all high performing countries, entry into teacher training programs typically takes place after the candidate has completed the schooling cycle or the equivalent of Grade 12. The length of the teacher preparation program afterwards does vary even across high performers. Ingersoll (2007) A Comparative Study of Teacher Preparation and Qualifications in Six Nations in a comparative study across six countries notes the variation in terms of the entry point into teacher preparation phase is non-existent, with all requiring a high school diploma. However, thereafter entry into the teaching

profession varies considerably from just having a High School diploma for elementary grade teaching in China, to an Associate's Degree in Singapore for teaching in elementary classes, to a full Bachelor's Degree required in Canada, Japan, Korea, Thailand and the US to become eligible to be a teacher. *Chile's requirements are no different from many other high performing countries.*

50. An important aspect of entry into teacher preparation programs focuses on the rigorous nature of the pre-entry processes. High performing countries like Japan, Korea, and Singapore employ high stakes entry procedures to limit the number of entrants into teacher preparation programs. Though there are no standardized assessments, Finland employs very strict standards for entry into teacher preparation programs. This is achieved in two ways – control in the number of institutions authorized to offer teacher preparation and being very selective in terms of the candidates who qualify from the student pool. In the case of Finland and Singapore, only a single institution in each country is authorized to impart teacher preparation courses⁴⁰. In comparison, in Chile, the US and Australia, the growth in institutions offering teacher preparation programs has been enormous. There are over 1500 institutions involved in teacher preparation in the US and in a recent study by the National Council for Teacher Quality (2013), it has been shown that except for a handful, the rest perform very poorly. In Chile, there has been an almost *uncontrolled* growth in the number of institutions offering teacher preparation programs and therefore ensuring quality standards becomes difficult. *Chile will need to revisit the requirements for the establishment and functioning of institutions that offer teacher preparation programs. In particular, given its size and the size of its student body, the number of students in teacher preparation programs seems disproportionate and the number of teacher training institutions too many for effective quality control mechanisms to function.*

51. Student selection into teacher preparation programs also varies considerably across countries and this can be typically classified as being through low, medium or high stakes channels. The top performers in PISA – Singapore, Japan, Korea, and Hong Kong tend to have medium to high stakes entrance requirements. Control at this stage is perhaps a major determinant of teacher quality in the classrooms⁴¹. In many of the best performing countries, teachers are drawn from the upper end of the distribution on the basis of their performance in college or high school – sometimes through a nationwide stand-alone test or simply based on performance at the school level. For example, in Finland and Singapore, teacher trainees are selected from the top third of the graduating class. In the US, teachers tend to be recruited from the top half of the distribution. Singapore offers top performers scholarships to complete their teacher education programs but with the understanding that these students will then teach sign a bond for a specified period of time⁴² to teach in public schools. In Chile, entry to teacher education programs is

⁴⁰ Both of these countries are small compared to Chile with about a third of the population. Even allowing for this difference, the fact that Chile has over a 100 institutions offering teacher preparation programs while Finland and Singapore have one each, suggests that there might need to be stricter guidelines to establish and run such institutions.

⁴¹ As Stewart (2012) in *A World Class Education – Learning from International Models of Excellence and Innovation* illustrates most countries in the world fail to limit entry into teacher preparation programs thereby creating an oversupply of poorly qualified teachers and in the process devaluing the entire profession.

⁴² Much has been said about the “*prestige value*” of being a teacher in some of these countries. While this is certainly important, *prestige* is defined by Hargreaves (2009) as *influence, reputation, or popular esteem derived from characteristics, achievements, associations*, while *status* is defined as *position or standing in society, rank, profession, relative importance*. However, this varies tremendously across countries and systems. While intrinsic motivation drives individuals to become teachers, it is important to buttress this intrinsic motivation with appropriately designed policies and incentives, or extrinsic factors.

contingent up students passing the university entrance examination and getting enrolled in a teacher education program. Thus, entry is based on a high stakes assessment. However, despite the high stakes process, most students who go into teacher preparation programs are selected from the tail-end of the distribution of those who appeared for the university entrance assessment. In a recent book, Bruns and Luque (2014) illustrate these entry level differences by reviewing scores from university entrance tests in Chile – and show that the average scores of students in Medicine, Engineering, Law and Teaching Schools are 745, 700, 645, and 505 respectively⁴³. Of course, comparing Chile with Singapore would not be wholly appropriate since in the latter all teacher-training is carried out through a single institution, the *National Institute of Education*, while Chile has over 90 institutions involved in the preparation of teachers⁴⁴, nevertheless it is important to identify the approach since it has policy implications for Chile – should entry into teacher programs be made more restrictive? *Chile will need to work on two fronts – raising the quality of teacher preparation programs and to ensure they can develop mechanisms by which the best students could be attracted to the teaching profession. While there are several policy levers that could be tweaked, the use of incentives to attract high performing students to the teaching profession through promises of civil service status or higher pay or promises of future scholarships conditional upon completion of a period of teaching in a public school could be viable options. Closing down existing but poorly performing programs is more complicated, though this may be a step that needs to be considered as well. These are both targeting upstream mechanisms for quality control. The Prueba Inicia offers a simpler downstream filtering mechanism but this would require new legislation to make the assessment mandatory and requiring all teachers to undertake the assessment, in addition to technical and administrative changes that will need to be undertaken to strengthen the quality of the assessment.*

Exit Requirements

52. Exit requirements from teacher preparation programs also vary significantly from country to country. There are two main issues to consider in terms of exit requirements and these are discussed in detail below:

- (i) Is there a stand-alone licensing/certification procedure or are licensing/certification procedures built into the teacher preparation program?
- (ii) Is teaching practice a requirement?

Licensing and Certification

53. Singapore, Korea, Japan, Finland, and many other high performers do not have stand-alone licensing or certification examinations. They require teacher trainees to complete their teacher preparation programs from an accredited institution, and base their entry into the teaching profession on a mix of course work requirements, passing of university tests and examinations, and maintaining an appropriate Grade Point Average. Of the OECD members, stand-alone teacher licensing and certification assessments outside of the institutions where they undertook their teacher preparation programs are held only in the US and the UK. In fact, such assessments are much more typical of practices in the developing part of the world, where the competition to enter the teaching profession is high and where

⁴³ Similarly, they also that at the University of Sao Paulo, students applying for law or engineering, and students applying for medicine scored 36 percent and 50 percent higher than students who were applying for teacher preparation programs.

⁴⁴ Of these institutes 52 are within university settings – 15 in the public domain, 37 in the private domain, 18 Professional Institutes (IPs) and 21 Technical Education Centers (CFTs). Only the IPs and CFTs created before March 10, 1990 are allowed to offer degrees in ECD and Elementary Education, the others are not.

quality assurance at the higher education level is weak. Most other countries accredit the teacher preparation program/institution, and completion of that automatically earns the student a license or a certification to teach⁴⁵. *In the case of Chile, the Prueba Inicia allows for the introduction of a formal system of ex-post licensing and certification. However, as noted earlier this will require legislative, administrative and technical changes before this can be achieved.*

Importance of Teaching Practice

54. Another area where countries differ broadly in terms of requirements is teaching practice. An emerging area in teacher preparation is the understanding that in high performing countries a substantial proportion of the time is spent on actual teaching practice. It is now recognized that high quality teacher practice, in settings that mimic real life teaching or are actually real-life teaching situations, are particularly important in enhancing student learning and development (Birch and Ladd, 1998). Studies of teaching practice illustrates three main channels through which these programs function – providing an encouraging and motivating environment in which to learn, helping trainee teachers understand the importance of classroom management, organization, and interaction skills, and finally, supporting prospective teachers on content, instructional and curricular areas of teaching. Doug Lemov (2010) in his book entitled *Teach Like a Champion – 49 Techniques that Put Students on the Path to College* – illustrates that teaching is an art and that the best teachers are those that not only know their content areas as expected, but are also those teachers that are best at interacting and communicating with the children in their classrooms. By observing the best teachers and classifying some of their techniques, he identifies a set of techniques that make these teachers exceptional⁴⁶. Emotional support in early grades is particularly important to ensure that students understand the process of learning and the realization that learning involves learning to fail and that practice is an exceptionally important part of any learning process. Finally, on the importance of instructional support, teaching practice helps trainees understand how to engage with students on a number of different levels and to help them work through their mistakes and celebrate their successes. Once again, this is particularly important in early grades as this helps form the attitude towards learning later on in life. See Box 6 below.

55. As we noted earlier, there is enormous variation in the lengths of teacher preparation programs. For example, one or two year teacher education programs after high school completion in China and Singapore for teachers who teach elementary school, to 3 or 4 year Bachelor degrees like in the UK and US followed by specific programs aimed at strengthening pedagogical knowledge and skills, or as in Finland where all teachers except for pre-school teachers have to possess a two years Master's degree after the completion of a three-year Bachelor's program⁴⁷. Beyond this variation, programs also vary in the proportion of the time that teacher preparation programs allocate to *actual teaching practice*. The duration of teaching practice varies from as low as 6 weeks in Australia, to 6 months in some European countries, to a year in Japan, and to 18 months in some of the Scandinavian countries. There are even

⁴⁵ Licensing and Certification in some countries is for a fixed period of time, and other countries it is for life - that is, once certified you do not have to have to be re-certified as is the case for some professions. Countries have also developed alternatives paths to becoming a Certified Teacher. There are procedures for teachers to obtain Advanced Certification or obtain the role of a master teacher. For example, Government's may decide that the shortage of STEM teachers is crippling and introduce mechanisms to induct STEM teachers from outside of Teacher Preparation programs, as long as they have been trained in the relevant subject.

⁴⁶ The book also provides an innovative set of videos that can be viewed at www.wiley.com/go/lemovideos. These videos illustrate the teaching techniques that Lemov presents as critical to high quality teaching.

⁴⁷ Finnish teachers aiming to teach in primary school need to major in Education and minor in two curriculum areas, while secondary school teachers be content specialists and major in the subject they will teach. The addition year or two is spent on mastering their skills either together with their coursework or afterwards at the end of which they obtain a master's degree

variations within the teaching practice requirement – in terms of whether this is supervised or not supervised, whether paired with an experienced teacher or not, or whether these teachers receive coaching support or no coaching support, etc. *Teaching practice does not seem to be a major part of the teacher preparation programs currently in place in Chile. While some teacher preparation programs clearly place emphasis on teacher practice, this is not a systematic feature of teacher preparation program in Chile and needs to be strengthened considerably.*

Box 6
Un Buen Comienzo⁴⁸

A project known as the Un Buen Comienzo supported by The Fundación Educacional Oportunidad, Harvard University and Universidad Diego Portales in Chile focused on improving teaching practices in urban schools in Santiago serving students from disadvantaged backgrounds in pre-kinder and kinder classrooms. Schools were randomly assigned across three groups – a full UBC module (Intervention 1), a partial UBC module in which books were distributed and a self-care workshop provided (Intervention 2) and a Control group. Pre-Kinder children were assessed prior to entering and once again at the end of the first year on numerous measures including language and literacy, socio-emotional skills, health, and attendance developed using the Classroom Assessment Scoring System (CLASS). Classroom interactions were videotaped and scored using a validated measure of classroom quality. The results illustrate that Chilean teachers performed poorly in terms of classroom interactions compared to their peers in the US. The UBC evaluation illustrates that Chilean teachers are able to quickly assimilate techniques for emotional support and classroom management once they have understood these practices. However, an area where they struggle is to bring instructional support to their classrooms. The UBC evaluation suggests improvements can be achieved by strengthening initial teacher training but also by supporting current teachers through the provision of in-service teacher training aimed at classroom organization and support to strengthening methods of student interaction.

Coaching and Mentoring

56. While teaching practice is important, working as an apprentice under a more seasoned teacher has been found to have very positive results. Pasi Sahlberg, a great spokesperson for the Finnish model of education refers to some misconceptions in this new world of high stakes accountability. In a popular article, Sahlberg in a thought experiment asks what would happen if teachers from Finland were relocated to Indiana, and likewise, teachers from Indiana were relocated to Finland. He believes that the Indiana teachers would thrive in Finland, and the Finnish teachers would simply taper down to the average teacher effectiveness level in Indiana. He says this based on the fact that he believes the systems in place believe in the collective unit of the school and that it is not individual teachers who make a school good or weak. The role of coaching and mentoring is a regular feature of most schooling systems, but in the best systems, this role is formalized in the early years for entry level teachers and fully supported by the more senior teachers in the group.

⁴⁸ Treviño, E. , Yoshikawa, H. , Leyva, D. , Snow, C. , Barata, M. , Weiland, C. , Arbour, M. , Rolla, A. and Toledo, G. , 2012-04-22 "Teacher practices and learning improvement in Chilean preschool classrooms" Paper presented at the annual meeting of the 56th Annual Conference of the Comparative and International Education Society, Caribe Hilton, San Juan, Puerto Rico

Performance Based Assessment

57. **As teacher practice gains importance, it is important to recognize that systematic assessment of teaching practice is more complicated than say the assessment of content knowledge.** The typical program that prepares teachers uses a wide variety of instruments to assess their preparedness for teaching. These programs typically have built into them a set of assessments that are linked closely to the standards and curriculum. Such assessments measure content and pedagogical knowledge. Comparing the assessments of the Prueba Inicia with the Praxis series of assessments, shows a considerable degree of similarities. However, increasingly the assessment of pedagogical *practice* is also being measured and there are serious concerns regarding validation in such studies. An area of assessments that is increasingly gaining popularity in teacher preparation programs around the world is that of *Performance Based Assessments* (PBAs). In this approach, in addition to the assessment of basic skills and content knowledge, as done through the Inicia, effort is also made to adequately assess classroom performance of trainees. PBAs employ more realistic or authentic settings to assess candidate performance in classrooms. Finland, Singapore, Japan and other countries where teaching practice is emphasized have in place assessments of such practice. The challenge is to ensure that such assessments meet the necessary psychometric properties needed to permit meaningful inferences to be drawn from such assessments. Most importantly, the reliability and validity of the assessments are brought into question even when the assessments are authentic, meet standards, and are well implemented. Reliability concerns exist due to the fact that experts are needed to score such assessments. Experts would typically use a set of rubrics to assess the performance of a trainee. However, even with a set of rubrics, a certain amount of subjectivity or bias is introduced into the system and reliability is called into question. Using more than one assessor or evaluator to assess performance might be one way of addressing biases introduced by particular evaluators. Performance based assessments also have validity concerns. In an assessment of this nature internal validity refers to the ability of a particular item to measure the specific skill or standard that it aims to measure, while external validity refers to the assessments ability to use the response and generalize the students' ability over the domain or knowledge across a particular area. Performance based assessments require the student to actually perform specific tasks as a way of demonstrating the set of skills needed to be a teacher. They demand far more of student teachers than memorization of facts and principles, but require that the student has studied, understood and is able to apply what they have learned in real life settings. There are broadly four categories of performance based assessments – (i) observation based assessments, (ii) performance of tasks on-demand, (iii) child case studies and (iv) teacher portfolios. Each of these has their strengths and weaknesses, though in this report we will not review these in detail.

*Curricular Design*⁴⁹

58. **A key element that seems to differentiate high performing countries from others is the manner in which the curriculum is covered and the importance given to content knowledge**⁵⁰. As a general rule, in almost all countries, content knowledge is given greater importance for teachers who

⁴⁹ In this section, I focus largely on one country and one subject. Clearly, curriculum across subjects is very difficult to compare. However, a lot has been said about the success referred to as “Singaporean Mathematics” and so in this section I look at how this has been designed to be so effective.

⁵⁰ In a study by Michigan State University, researchers find that countries that perform well in Mathematics (in international standardized assessments) such as Taiwan and Singapore are different from the US in a fundamental manner. They identify that math teachers are better prepared because their math training as high school students tended to be stronger, and because teacher preparation programs are very selective and attractive given the excellent compensation package including pay, benefits and tenure associated with teaching jobs in these countries.

teach in higher grades and pedagogical skills are emphasized for those teaching in earlier grades. Curricular design is a key element of the difference across countries. An often cited reason as to why Singaporean children outperform by a large measure their American counterparts in assessments of mathematics seems to be in part because of how the Singapore curriculum approaches the subject compared to counties and states in the US. In a study comparing Singapore and US Math Curricula, researchers find that on average, per grade, Singapore covers far fewer topics than do schools in the US. The table below illustrates that the average number of topics covered in Singapore per grade is about 15, whereas the seven studied American states range from 18 topics per grade in North Carolina to 39 topics per grade in Florida. It is interesting to observe that in both North Carolina (and Texas) which are in some ways closer to the Singapore model, NAEP Math scores have improved thereby suggesting that a well-defined curriculum focus is perhaps an important determinant of test performance.

59. While the issue of curriculum is well beyond the scope of this study, the importance of this in determining student learning and outcomes cannot be over-emphasized. Numerous studies have suggested that the reason for better performance of Singaporean or Korean students in Math is because of more highly qualified teachers, and this in turn has been linked not only to training that they receive in teacher preparation programs, but the emphasis they receive in Math during their own schooling. While most schooling systems do refer to programs targeted at individual students and the importance of learning at a pace which the student finds acceptable, in most countries this fails in practice – *children are left behind*. The top performing countries (albeit smaller in size) focus on achieving this objective. Recognizing that some children may have more difficulty in learning math, Singapore allows for a two track system – in which highly qualified teachers are brought to help students with difficulties learning particular subjects and helping them achieve *their learning goals mandated by the state* but at a slower pace than others. Such a framework allows for completion of all necessary topics, but at a pace more amenable to individual student needs. *Ramirez (2004) suggests that the poor performance of Chilean students on assessments such as the TIMSS is in part explained by the fact that curricular coverage is weak*. The distinction being made here is that while the US has an excellent curricular coverage of mathematics topics, though implemented poorly, in the case of Chile, prior to the revisions in 2002 of the Math curricula, the coverage was poor and hence resulted in the outcomes seen. A point to note however, has been the continued poor performance of Chile in these international assessments such as PISA and TIMSS.

Conclusions and Policy Options

60. Teaching is a complex task. A typical teacher needs to develop and maintain relationships, and build trust and confidence with numerous different clients. To complicate matters these clients tend to be of varying abilities, varied interests, typically have many different objectives, concerns and goals, and in particular, a set of these clients are at a stage of their lives where the work in school does seem irrelevant to their daily lives. Of course, while a teacher is dealing with enormous degree of complexity, they also have to simultaneously try and ensure that children in their classrooms are learning, and that what they are learning is relevant and useful in the future. Capturing and evaluating the skills needed to succeed in such a complex process through any one instrument is near impossible – many different levels and layers of assessments are needed.

61. **In August 2012, the previous Government considered making the Prueba Inicia a mandatory exercise for all teachers as they exit teacher training institutions.** Given the enormity of the reforms being considered by the Government, each element will have to be considered individually and efforts to make them into law taken by the Government. In an interview in May 2014, the Under Secretary of Education was asked whether the Prueba Inicia would become mandatory or obligatory, and her response was that it was not clear at this point whether the government would move on making this law. This analysis shows that while the test has some weaknesses these may well be improved by introducing more flexibility and strengthening programs that create incentives for professional development and accreditation.

62. The Prueba Inicia is an exit examination that assesses the content knowledge and pedagogical skills of teacher trainees as they leave teacher preparation programs. Currently, the Prueba Inicia is a voluntary exam, and as such is not high stakes assessment given its voluntary nature. However, a key debate currently taking place in Chile is whether or not to make the Prueba Inicia a mandatory assessment for trainees exiting teacher preparation programs and eventually use this instead as a credentialing tool for purposes of teacher licensing. The answer to this question is not straightforward. What do we know?

63. The following are stylized facts:

Table 4
Stylized Facts on Teacher Preparation

SNo	High Performing Countries	Chile
1	Do not typically have standalone licensure examinations for those entering the teaching profession	Chile at the moment also does not have such a system
2	Upstream quality assurance instruments are used including: <ul style="list-style-type: none"> - institutional accreditation - well defined curricular structures - limit the number of institutions involved in teacher preparation and - the number of students enrolled in them 	<ul style="list-style-type: none"> - Upstream quality assurance mechanisms weak - curricular structures have been recently improved and are of high quality, - however the number of institutions has grown very quickly and - with it the number of students
3	Comparing the PRUEBA INICIA with other similar assessments such as the PRAXIS series <ul style="list-style-type: none"> - Tests assess content and skill knowledge - Praxis tests are very well designed and 	<p>The Prueba Inicia is similar in design and structure to the PRAXIS tests</p> <ul style="list-style-type: none"> - Tests assess content and skills knowledge - Prueba tests are well designed and

	<p>developed</p> <ul style="list-style-type: none"> - Documentation is excellent - 	<p>developed in general, though there are weaknesses that can be addressed at the item level</p> <ul style="list-style-type: none"> - Documentation on the Prueba tests can be strengthened in keeping with best practices elsewhere. The documentation on the actual tests can be improved considerably to make them look more professional and make them more useful to examinees.
4	<p>Psychometric Properties of the Praxis is well researched and presented. Given the vast number of states which use the assessments and given the large number of candidates, and hence data on how the assessments perform, the properties of the Praxis have been well researched.</p>	<p>Although the overall development of the Prueba Inicia assessments also seem to have been done well, there are some key issues with the versions that we analyzed. In particular, the item level analysis illustrated a surprisingly large number of items that would need to be modified or dropped and new ones introduced to make the assessments more reliable. The two main concerns that would need to be addressed include reliability issues with particular items and the manner in which the levels of Insufficient, Sufficient and Outstanding can be finalized and the intervals between them be determined.</p>
5	<p>Assessment of teaching practice is weak in the current series</p>	<p>Assessment of teaching practice is also weak in the Prueba. However, this is increasingly being identified as an important step forward.</p>

64. The PRUEBA INICIA is just one tool among many which will be needed to improved the overall quality of teaching, teachers and of teacher preparation programs in Chile. In this section we review a few critical policy questions that are of relevance to this discussion.

65. These include:

- (i) Does Chile need an assessment of the type of Prueba Inicia?
- (ii) If no, what are the alternative mechanisms by which teacher quality can be enhanced?
- (iii) And, if yes, what are the issues that need to be addressed in the current form of the Prueba Inicia for it to have greater impact?

We conclude this part of the report by tackling each one of these briefly.

66. Does Chile need the Prueba Inicia? While there are many technical issues that can be raised about the specific nature of each test or assessment, the primary policy question confronting Chile is whether or not a Prueba type assessment is needed? Under the previous government there was a very keen desire to introduce legislation to make the Prueba Inicia mandatory. Furthermore, there were proposed measures to link performance on the Prueba Inicia to initial salary levels of newly recruited teachers. While such high stakes accountability measures have been introduced in some countries, purely from the view point of public policy it could be argued that they are yet to demonstrate the desired effects

in terms of improved teaching quality. What Chile needs is a mechanism by which the country can ensure that all teachers will be fully equipped to teach their classes before they are placed in front of students. While this can be done by putting in place a high stakes teacher licensing system, there might be more meaningful ways by which these changes could be brought about. If there were absolutely no means by which upstream quality control processes could be introduced, the Government would have no other option but to introduce a filtering mechanism downstream. However, Chile has institutions that function and can be strengthened.

67. There are other reasons why an stand-alone Prueba Inicia may not be the most useful model. As we have noted, a Meckes et al illustrate teachers have shown a lukewarm response to the assessment – and the reasons typically for such a response is that such assessments have very little value within classroom settings. These measures of teachers’ competency as noted earlier lack authenticity and predictive validity, say when compared to performance based measures which based on the assessment of teaching practice is a better predictor of teaching ability. Given the scant evidence surrounding teacher credentialing tests and the lack of evidence on predictive validity of these tests in identifying effective teaching, we would need to reconsider overhauling the entire approach.

68. The lowest hanging fruits in terms of upstream mechanisms revolves around incentives to individual students. Combined action of raising the minimum qualification levels for entry into teacher preparation programs, while simultaneously offering scholarships for further studies in subject matter area upon the completion of four or five years of teaching in public schools, could achieve twin goals at the same time: (i) could compel institutions offering teacher preparation programs to either find students with very high PSU scores or it would compel a number of the institutions to close doors due to a lack of students, (ii) secondly, it would put in place mechanisms to attract top students initially into the teaching profession which at present seems a challenge. Many countries that are eager to try and attract their best students into teaching offer targeted incentives to attract students who would otherwise choose engineering or medicine or law as their career choices⁵¹.

69. Other upstream quality assurance measures are more difficult to achieve in the short term. However, Chile has in place a quality assurance system and systems for institutional accreditation and as these are strengthened, the need for a standalone licensing/certification assessment will diminish.

70. Alternatively, if the Government believes that even the most basic upstream quality control measures cannot be put into place in the medium terms, then downstream processes have to be brought into play and the Prueba Inicia should be mandated for all students exiting from teacher training or preparation programs. This would involve strengthening of the Prueba Inicia instrument to ensure reliability needed for a high stakes assessment. Furthermore, given that the Prueba Inicia does not include an assessment of teaching performance, it would be important to introduce a system of performance based assessments to focus in parallel more directly on the skills needed to become an effective teacher.

⁵¹ Singapore, Finland, Korea, Taiwan all employ similar policies.

DETAILED PSYCHOMETRIC ASSESSMENT

METHODOLOGICAL ISSUES

Analysis

71. The psychometric analysis is applied to the twelve tests that compose the *INICÍA* battery. These include the following: (i) one written test based on the thesis of the student (*PCE-INICÍA*), (ii) three tests of pedagogical knowledge (*PCP-Básica*, *PCP-Media*, and *PCP-Parvulia*), (iii) two subject areas tests in preschool and primary education (*PCD-Básica* and *PCD-Parvularia*), and (iv) six subject area tests in secondary education: Language (Spanish) (*PCD-Lenguaje*), Math (*PCD-Matemática*), Biology (*PCD-Biología*), Chemistry (*PCD-Química*), Physics (*PCD-Física*), and History (*PCD-Historia*). The sub-scores and the total score of the first one was re-coded for the item analysis. The basic statistics of the sets are collected in Table 4.

Table 4. Basic statistics of the components of the *INICÍA*

Set ¹	PCE	PCP			PCD							
Test ²	INICÍA	Bas	Med	Par	Bas	Par	Len	Mat	Bio	Qui	Fis	His
numerus ³	1,824	669	754	295	663	289	80	179	80	43	54	131
max. score	36	50	50	50	80	60	60	60	60	60	60	60
reliability ⁴	0.64	0.66	0.72	0.68	0.82	0.69	0.71	0.88	0.77	0.80	0.91	0.74

1) PCE = Prueba de Comunicación Escrita, PCP = Prueba de Conocimientos Pedagógicos, PCD = Prueba de Conocimientos Disciplinarios

2) Bas = Basica, Med = Media, Par = Parvulos, Len = Lenguaje, Mat = Matematica, Bio = Biología, Qui = Química, Fis = Física, His = Historia

3) Combined version A + B

4) Reliability of Theta as the mean of versions A and B estimated after equating the test scores

Psychometric Concepts And Methods of Analyses

72. The psychometric properties of the tests are analysed using two strategies: (i) by using modern test theory or Item Response Theory⁵² and (ii) the Classical test theory⁵³. The general and specific issues are handled in the following sections.

IRT-Modelling and Classical Test Theory

73. The main disadvantage of the CTT is that the statistics are always bound to the sample. Review the *INICÍA* assessments, shows that the number of *cases* in some of the datasets seems limited or very sparse. For example, there are only 43 cases in the *PCD-Química*, 54 cases in the *PCD-Física*, and 80 cases or test-takers in the *PCD-Biología* and *PCD-Lenguaje*. When sample sizes are small, the estimates of items parameters such as *difficulty*, *discrimination power*, and *guess* may be unstable. Secondly, it limits analysis to the one-parameter IRT model, that is, Rasch modelling to calibrate the difficulty levels of the items in the same scale and further to equate⁵⁴ test scores over the subjects and tests. CTT, including statistics such as *item discrimination*, the *percentage of correct answers*, and *reliability*, is used

⁵² Also referred to as IRT-modelling. Refer to Rasch (1960), Lord and Novick (1968), Lord (1980) and Hambleton (1993) for further details.

⁵³ Also referred to as CTT. For a detailed exposition, please refer to Gulliksen (1950), (1997).

⁵⁴ For further discussion on equating please refer to Béguin (2000).

mainly in the deeper analysis of the items, that is, distractor analysis. Given the technical nature of this work, the following section introduces technical terms, statistics, and practices used in this report⁵⁵.

Specifics of IRT/Rasch Modelling

74. In order to understand the psychometric part of the report, we focus on three concepts within IRT modelling and their relations. These include: item difficulty β (*Beta*) which is essential in calibrating the item difficulties into the same scale, latent ability θ (*Theta*) which is essential in test score equating, and linking procedure between the tests which is essential in combining β and θ . Both β and θ follow a standardized Normal distribution.

75. When the tests are not strictly parallel, the final scores of the eleven tests are not comparable without proper transformations based on the calibration of the items into the same scale. This means that 30 points in one test is not necessarily comparable with 30 points from another test even though the maximum values in the score would be the same. The reason is that the difficulty levels of the tests may be different. From this point of view, the IRT modelling and related test score equating is the only credible way to compare test scores. The specific advantage of IRT modelling is that the latent ability level of a learner (θ) and difficulty level of an item (β) are, first, not dependent of the sample, and, second, they are identical when certain preconditions are met⁵⁶. Hence, the latent ability for each pupil can be determined in the same metric for every test as far as there are linking items connecting the versions. Now, practically all the test-takers ($n = 1824$) did the written part of the test, *PCE-INICÍA*. The six criteria for assessing these (Spelling, Text Cohension, Vocabulary, Thesis, Structure, and Argumentation), that is, “items” on the written test were used as the linking items for the rest of the tests. Technically, the six items on the written test were added to the other tests to be the linking items. The original scoring in *PCE-INICÍA*, however, was amended to fit the IRT practices: the scoring system should be made of whole numbers and all the categories should be observed. Hence, in Spelling, for example, the original scoring and frequencies were as follows:

Original Score	1	1.25	1.50	1.75	2	2.25	2.50	2.75	3	3.25	3.50	3.75	4	Total
Frequency	1001	1	266	4	261	0	132	1	132	0	20	0	6	1,824

76. However, as the frequency table above illustrates the real scoring is 1, 1.5, 2, 2.5 and so on up to 4. By lowering the 0.25s systematically into the lower category and 0.75s to the upper category, this was transformed into the following systemic:

Reduced score	Final score	Frequency
1	0	1002
1.5	1	270
2	2	261
2.5	3	133
3	4	132
3.5	5	20
4	6	6

⁵⁵ For a deeper treatment of these issues please refer to Gulliksen (1950) or Metsämuuronen (2013).

⁵⁶ Refer to Wright (1968) and Metsämuuronen (2013).

77. The difficulty parameters (β) of the PCE were estimated first with 1,824 applicants. After that, the six PCE items were added to each set of tests for the applicants as the linking set of items. The estimated values of the item difficulty parameter of PCE were fixed and the item difficulties for all the other items were freely estimated. Then, the ability level of the average student would correspond with the latent ability of round $\theta = 0.00$ and the average items difficulty would be round $\beta = 0.00$. In what follows in test score equating, the Theta value refers to a certain test score; the test scores are not comparable over the tests but the latent ability levels (Thetas) are comparable across tests.

78. The estimation was run with OPLM program⁵⁷. Equating the test scores with IRT modeling was administered with the following principles and practices. A brief technical description of the equation process follows⁵⁸:

- i. Define the structure of the test so that the linking items are connecting the tests to each other. Because values of the difficulty parameter of the linking items are exactly the same in each version the difficulty levels of all other items are calibrated into the same scale as the linking items are.
- ii. Use *Conditional Maximum Likelihood* (CML) procedures to estimate the difficulty level (β parameter) for each item.
- iii. Use *Marginal Maximum Likelihood* (MML) procedures to estimate the distribution of each student's latent ability (θ parameter) in each version.
- iv. Estimate the θ parameter of the scores of each version using the means and deviations of the distributions of β and θ . This results in a unique latent value, however measured in a common scale, for each observed value of the scores in all tests.

Specifics of Classical Test Theory and Related Indicators

79. As in the IRT modelling, in CTT as well item parameters are of specific interest. In CTT, the parameters, such as the item difficulty (estimated by the percent of correct answers) and discrimination power (estimated by the item total correlation, R_{it} , and item-rest correlation, R_{ir} ⁵⁹) are, however, sample dependent. Of indicators of item discrimination, $R_{it} > R_{ir}$ by mathematical construct. Metsämuuronen

⁵⁷ Please refer to Verhelst, Glas and Verstralen (1995).

⁵⁸ For a more exhaustive treatment of the same, please refer to Béguin(2000) pages 17–36.

⁵⁹ The assessment will have to discriminate between high and low performers, implying that truly high performers should have a higher probability of responding correctly to any item, as opposed to having an item in an assessment where poorer performers have a higher probability of correctly answering the item. Of course, if the latter happens, construct validity might be called into question. The item-total correlation ranges from -1.00 to 1.00. An item is said to discriminate well between high-performing and low-performing participants, when the value of the item-total correlation is high and positive. If the item-total correlation is negative, low-performing participants have a higher probability of getting items correct. Items that are not capable of discriminating well have item-total correlation values closer to zero and the both high performers and low performers, regardless of their total assessment scores are equally likely to answer an item correctly. There are three issues which need to be considered – (i) when items are scored or weighted differently, (ii) when the assessment has too few items, and (iii) when the same size is small. In assessments where some items are weighted higher, for example, scored as 0 or 10, and others are scored as 0 or 1, then students who score the 0/10 item correctly immediately score 10 more points in their total. Even when each item is weighted or scored in a similar fashion, when the number of items is few, then each has a bigger contribution to the total score. And, finally, the stability of correlation coefficients as noted above comes into question when sample sizes are small as was noted for some of the INICIA assessments. The Item Rest Correlation is a way of addressing some of these issues and in this the correlation coefficient does not include the contribution of the item to the correlation coefficient.

(2013) shows that both Rit and Rir always underestimate item discrimination and Rir underestimates more than Rit . Hence, in this report mainly Rit is used as the indicator for the item discrimination.

80. The classical boundary for acceptable item discrimination is $Rit \geq 0.20$. Item discrimination is typically maximized for those items of medium difficulty. When an item is very easy (or conversely, very difficult), it is rare to find an item with very high Rit . In that case, even a somewhat lower value of 0.18 to 0.19 could be acceptable. However, when the item is of medium difficulty level, it would be expected to have a much higher value of Rit . As a general rule, when $Rit < 0.20$, the item is considered to be poor from the point of view of discrimination, and only in exceptional situations⁶⁰ should the item should be selected into the final instrument.

81. When the item discrimination is negative, that is, $Rit < 0.00$, the item is considered *pathological*. This means that the weaker test-takers become more likely to obtain the correct answer compared with better performers. If item discrimination is notably below the zero, it could be because an incorrect *key* was provided and not the correct one. So, while the right answer might have been distractor A, the key identifies B as the correct distractor.

82. Two other indicators that have a specific role in evaluating the psychometric properties of INICIA are mentioned. Assessing individual distractors of the multiple choice (MC) questions involves looking at both Rir - and Rar values. Rir is the item-rest correlation as defined above. Rar is the correlation of the alternative (distractor) and the rest score. The items are flagged in three cases:

- i. If $Rar \geq Rir$, a distracter correlates as high as or higher with the test's rest score than the correct alternative,
- ii. $Rir \leq 0$, the correct alternative does not correlate or even correlates negatively with the test's rest score, and
- iii. $Rar \geq 10$, a distracter - test's rest score correlation is suspiciously high⁶¹.

83. One additional, technical, note of the connection of the item discrimination, item difficulty, and test reliability: The item difficulty and item discrimination, classically estimated by using the proportion of correct answers (p) and the item-total correlation ($\rho_{gX} = r_{it} = Rit$), are interrelated so that the item discrimination is the highest when the difficulty level is around 0.50. When knowing that the variance of the dichotomous item is strictly related with the item difficulty, that is, $\sigma^2 = p(1-p)$, the classical formula of *Alpha* reliability can also be expressed with these two indicators as follows:

$$rel = \frac{k}{k-1} \left[1 - \frac{\sum_{i=1}^k \sigma_i^2}{\left(\sum_{i=1}^k \rho_{gX} \sigma_i \right)^2} \right]$$

where k = number of items
 σ_i^2 = variance of the scores on item i
 σ_i = standard deviation of the scores on item i
 $\rho_{gX} = r_{it}$ = item-test correlation

84. Only two sources of information are needed for estimating the reliability of the MC-test: the item total discrimination (Rit)⁶² and the item variance (σ_i^2) or item difficulty (p). It is also noteworthy that the Alpha reliability is maximized when the sum of the elements $Rit \cdot \sigma_i$ is the highest. Knowing that the variance is the highest when the proportion of the correct answer is $p = 0.50$, it follows that it is best to select items with as high item discrimination as possible and medium difficulty level to maximize the reliability of a test.

⁶⁰ For example, when willing to ensure the validity of the test.

⁶¹ TIAPLUS software (Heuvelmans, 1998) calculates many other indicators for the individual item.

⁶² Note: not the item-rest correlation, Rir

PSYCHOMETRIC ANALYSES

Introduction

85. The datasets are analyzed in five ways.

86. Classical test theory and item analysis is used to acquire the discrimination power of the items and overall test reliability. When sample sizes are small⁶³, a two-parametric IRT model, which would produce item discrimination automatically, is not recommended. But by combining item difficulty with discrimination, it is possible to assess which of the items are poor or even pathological.

87. Classical statistics are used also in analyzing *why* the flagged items are poor or pathological. The aim here is to find distractors which lead items to be poor or pathological. The analysis can also help suggest whether the key was correct or not, also hints as to whether the key is correct or not, and whether there is pathologically high guessing with respect to this.

88. Mantel-Haenszel (MH) statistic is used also in analysis of Differential Item Functioning (DIF). If two groups of test takers with *equal ability levels* have systematically different probabilities of responding correctly to an item, Differential Item Functioning (DIF) is said to occur⁶⁴. A key issue in the determination of whether DIF exists or not, relates to the sample size concerns. In particular, whether there is sufficient sample size for the *Reference Group* and the *Focal Group* being compared and whether the sample size is large enough and with sufficient statistical power needed to identify DIF. Unfortunately, many data sets have too few test takers and whether DIF can be identified or not is also a function of the statistical formulation being used. The Mantel Haenszel is a preferred approach when sample sizes are smaller and make the use of IRT more difficult. The MH approach looks at differences across reference and focal groups, across the ability spectrum, for all test items individually. Once the groups have been classified and their responses correctly coded, the odds ratio for the groups is obtained based on the proportion of correct and incorrect answers for the two groups. The odds ratio varies in value from 0 to infinity, with odds ratio of 1 representing the point at which there is no DIF, and odds ratios between 0 and 1, and above 1, representing points where the Focal group outperforms the Reference group and vice versa.

89. One-parametric IRT modeling (Rasch modeling) is used to acquire the sample-free item serial difficulty. The aim of this is to calibrate the items over the tests into the same scale and to acquire comparable item difficulty levels. This is done by using OPLM software. Additionally, the standard IRT modeling allows for the graphical evaluation of the Item Characteristic Curve, and hence, flag the possibly pathologically behaving items. Here, however, roughly the same is done by using the distractor analysis (see point 4).

90. The sample-free item parameters of the IRT modeling are used in equating the test scores over the different tests. The aim of this is to estimate the latent ability needed for each test score in each test version. By doing this it is possible to evaluate whether the original test scores are comparable – or rather, *which* of the test scores are comparable. This is important in assessing whether the boundaries of “Outstanding”, “Sufficient”, and “Insufficient” are comparable over the tests. This is done by using OPLM software.

⁶³ For example, with 80 test-takers in PCD-Biologia, 54 in PCD-Fisica, and 43 in PCD-Quimica.

⁶⁴ An item is said to be biased if the underlying reasons for the DIF is not part of the test construct.

Restrictions of the Analysis

91. Although the psychometric analysis of the tests is quite thorough, there are two limitations. First, subject matters analysis is not included, and second, the issues of language are not tackled. In most assessments of this nature, the contents of the assessments are bound to “theoretical frameworks” which is based on a broad consensus of the nature of the subject being taught and the specific contents to be pursued at each grade level. Inclusion of this subject matter assessments is beyond the scope of this work and would require not only a thorough knowledge of the Chilean curriculum, how these are taught in schools and teacher education institutions, and one expert for each subject and level being assessed⁶⁵. This evaluation focuses more on the technical properties of the items and less on the broader links between national subject specific standards and the tests.

92. The second weakness in the assessment is language. Item development needs to be done very carefully and, in particular, the nuances introduced by different formations may result in items being designated as poor or even pathological. These language issues are very delicate and are not tackled in this report. The analysis helps flag some items, other experts both content knowledge and/or language experts will be needed to check each and every flagged item for such issues⁶⁶.

Results

Development and Implementation of the INICIA

93. Documentation of the INICIA, its overall objectives, test development methodologies, procedures during piloting, and the characteristics of the items and tests is done very well. These are thoroughly reported in four reports⁶⁷. The reports describe the procedures employed during the piloting phase, implementation and results of the pilot test processes, the samples used, the protocols, security and confidentiality measures; the technical aspects of the tests: instruments and composition of the content, difficulty levels and cognitive levels as well as the criteria for selecting items to include in the final assembly; the criteria for the preparation of the final test: the analysis of failure rates, difficulty index, discrimination index and analysis of incorrect options or distractors as well as the psychometric characteristics of each of the assembled items. The reports give proposals also for the criteria and guidelines for the process of correcting tests and a set of recommendations for future disciplinary processes tests and a detailed explanation of the elements that could be improved in the parallel process in future. This publicly available information is often considered as *best practice* in similar such assessments in other countries.

94. The impression that a lay reader gets when reviewing the documentation, is that this is professionally done⁶⁸, exhaustive, and helpful for the next round of test constructors. Sustainability is built into the process by involving relevant units of universities to do parts of the work.

⁶⁵ Although the support of subject experts was initially envisioned, this was for the purpose of reviewing the specific items that were determined to be poor performers or pathological and not to ensure that the tests were fully integrated with the curriculum and standards envisioned in the Chilean education system.

⁶⁶ Metsämuuronen notes (in a private conversation) that while evaluating thousands of Finnish items by evaluators, it was observed that small and delicate wording nuances caused some items to become poor- or pathological ones. In most cases, he noted that these nuances in Finnish were detected only after finding that the item is poor – *not before the pretest*.

⁶⁷ These reports are entitled Evaluación1 (2013); Evaluación2 (2013); Evaluación3 (2013) and Evaluación4 (2013).

⁶⁸ The reported procedures of the test assembly fulfill the criteria for professionally-done work. This includes the following often considered best practices: item writers were selected from a pool of experienced professionals, as

95. Two other issues may be worth noting. First, no documentation is found of the final testing, and the related procedures. This is problematic especially considering the effort that has gone into the documentation at the time of design. Hence, it is practically impossible to assess data management and analysis or scoring procedures of the final phase. Second, it seems that sample selection during piloting was probably not very successful. As a summative assessment, and given the high stakes nature of the test, the piloting should have been done in a more confidential manner. However, the piloting sample seems to have been compiled using volunteer students and teachers. Given the presence of many low-discriminating items – even pathologically low-discriminating ones – and given that many poor items were reportedly omitted from the tests, the continued low quality of these assessments might be because of pilot phase sampling.

96. There are many ways to assess whether a test and/or a set of tests measures what it aims to measure. That is, to assess the validity aspects of a single test and/or a whole set of tests. The first question that we would need to consider is what is the purpose and aim of INICÍA? It is important that the INICIA identifies and states the purpose of the test publicly and ensure that all candidates for the assessment are made aware of the purpose of the assessment and have a clear understanding of the knowledge/skills/abilities being measured. The INICÍA aims “to monitor the knowledge and skills of new graduates from pre-teacher training institutions”. The tests are designed to measure the knowledge dimension of the new graduates. However, it seems that the set of tests, used alone, gives a restricted picture of the skills of the graduates (Meckes, 2012)⁶⁹. However, the knowledge base of the teachers is important part of the professional work.

97. Obviously, there are several other dimensions than the knowledge base in good teaching. As an example of a theoretical model of a “good teacher”, Metsämuuronen & Metsämuuronen (2013a; 2013b) suggest – on the basis of a literature survey of the Finnish teacher educators and an empirical data from Nepal – a four-fold model of a “good teacher”. In this model, a good teaching comprises four elements: (a) Personality of the Teacher, (b) Pedagogical Skills, (c) Content Knowledge, and (d) Classroom Management. Of these, the skill of the classroom management is an obvious need for a teacher in the situation when the children are taught in big groups. Muijs and Reynolds (2005, 75) argue that classroom management distinguishes the effective from the ineffective teachers. Content knowledge and pedagogical skills are inevitably bound together. That is, even if graduates have high scores in a test measuring content knowledge, if they are unable to transfer this knowledge to their students, then content knowledge alone will not improve student learning. Conversely, however, if a teacher or a graduate has good pedagogical skills, (s)he may pick the content knowledge or content pedagogical skills from good materials and peers.

98. Teacher personality– including *inter alia* child-centered aspects like kindness, fairness, being easy to approach and ask questions, supportiveness, and personal attributes such as, calmness, self-confidence, self-efficacy, and promptness or systematic aspects are all critical factors that contribute to being a good teacher. In fact, for beginning teachers who also tend to be assigned to lower grades in many schooling systems, these personal attributes may be even more important than lower grades in content knowledge. Teacher personality assessments are gaining strong foothold globally and are an increasingly important aspect to consider. Many countries are now carrying out teacher personality

were the others who played their part as test assemblers, the Table of Specifications were prepared adequately, the relevant stakeholders were involved in the processes or at least they were informed of the processes, the item analysis is done by using proper and adequate practices, and the confidentiality was secured during the process.

⁶⁹ This is similar to asking a student of carpentry to describe the wood and tools used in their work, but not evaluating how (s)he actually applies this knowledge. The same way, the INICÍA does not seem to assess the skills of the graduate teachers adequately.

assessments as a way of ensuring that only teacher trainees who have the requisite communication and interpersonal skills to deal with a classroom full of students are being placed in classrooms. In England, the government has decided to ensure that teacher training programs will emulate practices in high performing countries, such as, Finland Korea, Singapore, etc. where trainee teachers spend a considerable amount of time under *supervised settings* in classrooms with real students. This is achieved through a formalized relationship between teacher training centers and specialized training schools where teacher trainees are supervised. The INICIA at this point in time does not assess performance in the classrooms even though there is some practice within training programs⁷⁰.

99. While understanding the restrictions of the *INICIA* examination from the validity point of view, it is still important to evaluate the validity of the tests as a part of Content Knowledge. Test validity is assessed or evaluated through four viewpoints: (i) in a general way as the Face validity, (ii) the structure of the tests as Construct validity, (iii) more specifically as the Content validity, and finally, (iv) how practical the examination is from the teacher's profession viewpoint as the Ecological validity.

Face Validity

100. This aspect of the assessment is typically used to obtain a "feel" for the assessment and the processes through which it is implemented⁷¹. The overall impression of the tests is that they are well prepared. Intuitively, because of the variation in text, graphical design, and use of tables and other such features, the assessments seem good from the test-takers viewpoint. The consistent structures of the assessments and the well described set of standards in the background documentation, make these assessments look systematic and well-thought out. Including a line on or short para on each assessment specifying the purpose of the test would be consistent with some of the best practice in the world. A key weakness in the assessment is that only multiple choice questions (MCQ) are used in the assessment in most parts of the assessment, except the assessment where an essay or a composition is needed. The inclusion of some open-ended questions or more demanding productive items and would enrich the tests as is practiced in the international standard in the student assessments (see, for example, Mullis & Martin, 2011, 6; PISA 2006). Hence, from the Face validity viewpoint, the tests are interesting, professional looking, and versatile though restricted to MCQs.

Construct Validity

101. According to the documents describing the development of the *INICIA*, the school curricula were used as the basis of the test structure of the *PCD-Básica* and *PCD-Parvularia* (Evaluación1, 2013, 5), the *PCD-Lenguaje* and *PCD-Historia* (Evaluación2, 2013, 10), and the *PCD-Media*, *PCD-Matemática*, *PCD-Biología*, *PCD-Física*, and *PCD-Química* (Evaluación3, 2013, 8). The final report 4 (Evaluación4, 2013, 11) supported the development of the *PCP-Parvularia*, *PCP-Básica*, and *PCP-Media*. In particular, the *Table of Specification* was prepared on the basis of *Estándares Orientadores para Egresados de Carreras de Pedagogía en Educación Básica, Parvularia o Media*. The structure of the written test is not reported but the division of six criteria for the assessment seems relevant. The structures of the tests are well-documented by the test developers, they are based on a relevant theoretical framework (school curricula), and the observed structures correspond with the intended ones (Tables 2–11). Hence, the structures of the tests seem valid.

102. Three additional notes of the structures of the tests may be worth giving. First, it seems evident that the aim in constructing the tests was to maximize the *validity over the reliability*⁷². This is based on

⁷⁰ For example, the PUCC program on basic teaching has 4 practical classes on teaching.

⁷¹ Though this is not taken very seriously in many settings, there is a benefit to having a short note or a paragraph.

⁷² This is also the approach adopted in Finland within the national student assessment (Metsämuuronen, 2009).

the fact that the number of items on some of the sub-tests is very sparse (see Tables 5 and 8). Only two or three items are selected to represent certain themes. This evidently leads to the Lord and Novick paradox, by maximizing the validity one minimizes the reliability and *vice versa* (Lord and Novick, 1968; Metsämuuronen, 2013). On one hand, measuring very accurately the wrong thing is usually less preferable than measuring the correct thing but with a less accurate manner. Maximizing the validity may be one reason why the reliabilities of the sub-tests of INICÍA are quite low ranging from $\alpha = 0.64 - 0.91$ (see Table 2.1).⁷³ On the other hand, the INICÍA tests are aimed for discriminating the students in a high stakes manner. When thinking the high stakes role of the INICÍA, only the reliability for the sub-test of *PCD-Física* ($\alpha = 0.91$) is high enough for discriminating the test scores (and ultimately the test-takers) from each other. The reliabilities of *PCE-INICÍA* ($\alpha = 0.64$), *PCP-Básica* ($\alpha = 0.66$), and *PCD-Parvularia* ($\alpha = 0.69$) are very low from this perspective.

2. A practical calculation may clarify the challenge of low reliability: Let us take the *PCP-Biología* as an example. The general, Classic, standard error of the measurement (S.E.M.) is estimated as follows: $\sigma_E = \sigma_X \sqrt{1 - \text{Rel}}$ where σ_X is the Standard Deviation of the total score and Rel is the reliability of the test. For *PCP-Biología*, $\sigma_X = 7.0788$ and the reliability, estimated by using the *Alpha* model, is $\alpha = 0.77$.

Hence, S.E.M. is $\sigma_E = 7.08\sqrt{1 - 0.77} = 3.39$. On the basis of this, one can estimate the error of a single score. At the final phase of the assessing of the students achievement level, the cut-offs for the “insufficient”, “sufficient”, and “outstanding” were set to 35 points (insufficient/sufficient) and 41 points (sufficient/outstanding) out of 60 points (see Table 16 in Section 5.3.5). The true ability of the test-taker with the score 35 (labeled as “insufficient”) could be also $35 + 3.39 = 38.39$ (that is, “strong sufficient”). On the other hand, the true ability of the test-taker with the score 41 points (labeled as “outstanding”) could be $41 - 3.39 = 37.61$ which is actually *lower* than the true score of the “insufficient” test-taker! Now, the order of these “insufficient” and “outstanding” test-takers would be opposite. That means that, in theory, in another day, measured with the same test, the “insufficient” test-taker would have been ranked as “outstanding” just by guessing correctly one item more and opposite: the “outstanding” test-taker would be labeled as “insufficient” just by being making an error with one item. It is very short way from the bottom to the top because of the low accuracy of the test.

103. The TIAPLUS software automatically performs the factor analysis to test whether the structure of the test is one-dimensional or not. In all cases with all versions, they seem to form *two* dimensions. These dimensions are not necessarily meaningful from the content-wise the same way as in the attitude scales; in the 0-1 matrix of an achievement test, the easy items correlate with each other and the difficult items correlate which each other and hence there tend to appear two or three factors when the test is compiled so that it includes multiple difficulty levels even though the content-wise structure would be different.

104. All tests seem to have two variants, A and B. In some tests, there are two different versions which are linked together with the anchoring items. The number of linking items is proper for the stable estimation of the items’ parameters over the versions. Some tests cleverly rotate exactly the same items so that their parameter values are not – most probably – affected by their position. Hence, the test versions can be taken as strict parallel tests (or, actually, the same test). While performing the IRT modelling, the items on the different versions are taken as the same item even though the position of the items may have a slight impact of the difficulty parameter.

Content Validity

⁷³ Another, related, reason is discussed in what follows; the main technical reason for the low reliabilities is in the low item discrimination. This can be explained only partly by the structure of the tests.

105. As noted above, according to the final reports of the test development (Evaluación1 – 4, 2013), the contents of the tests were based on either the national curricula (*PCD-Básica PCD-Parvularia, PCD-Lenguaje, PCD-Historia, PCD-Media, PCD-Matemática, PCD-Biología, PCD-Física, and PCD-Química*) or the Guiding Standards for Educational and Alumni Career in Basic Education, Early Childhood or *Media (PCP-Parvularia, PCP-Básica, and PCP-Media)*. There is no doubt that the contents of the tests are valid to measure the knowledge base of the beginning teachers. To critically evaluate the contents of the tests needs a large and experienced team with 11 or 12 subject specialists.

Ecological Validity

106. Under this section, we review to aspects of the assessments. The first focuses on the coverage of the cognitive domain or the depth of the tests, while the second reviews the overall transfer of the test results to real world teaching. From the ecological validity viewpoint, test depth seems versatile for assessing the cognitive processes of the student teacher. The final reports of the test development (Evaluación1 – 4, 2013) specify the structures of the tests anchored to Bloom's taxonomy of the cognitive domain (Bloom *et al.* 1956; Metfessel, Michael, & Kirsner, 1969).

107. In the simplified version – suitable for the national level testing – the original taxonomy can be reduced into four: (a) Knowledge or Recall, (b) Comprehension, (c) Application, and Higher skills. For example, in the Program for International Student Assessment (PISA) or the Trends in International Mathematics and Science Study (TIMSS), Comprehension and Application, seems to be combined (see PISA, 2006; 2009; TIMSS, 2007; 2009a; 2009b). In *INICÍA*, it seems that the Application and Higher skills⁷⁴ (“Analyze and the use of knowledge”) are combined. On the basis of the description of the contents of this category, it seems, however, that these items are geared toward higher skills though they are called “skill-related items”:

“Analyze and the use of knowledge: The graduate uses his/her disciplinary and pedagogical knowledge to analyze and evaluate information based on which should come to a conclusion. The graduate teacher is capable of hypothesizing and questions, clarify meanings or implicit information, generalizations, comparing evidence, critique concepts, models, actions, strategies, events or situations to make decisions. Importantly in these skill-related items (i) the question assessed knowledge is not explicit or direct, or (ii) requires to stake diverse knowledge to respond.”
(Evaluación4, 2013, 16)

108. In all the tests, the proportions of Knowledge, Comprehension, and Higher skills items were fixed to 30%, 40% and 30% respectively. Intuitively, the number of recall-type of items feels quite high. The international student assessment settings as PISA and TIMSS seem to be geared toward application rather than memorizing things.

109. Another perspective to the Ecological validity is obtained by asking how well the *INICÍA* test really reflects the graduate teacher's capability to teach the children in general and specifically of a certain subject. The question stays open. However, when compared to the Finnish reality, one might have to conclude that the possibilities are perhaps *limited*. In Finland, where the teachers' high quality is seen as one of the explaining factors for the high ranking in PISA studies (see, for example, Kansanen, 2003; Niemi, 2010; 2011; Niemi & Jakku-Sihvonen, 2006; 2011; Sahlberg 2011a, 2011b; Schleicher, 2011) the test like *INICÍA* would be taken too narrow for assessing the real capability of a young teacher. In Finland, the graduating teachers need to do *three months'* practical period in a real school under the teacher and a pedagogical expert (teachers educator) after which it is assessed whether they are capable or

74

competent to be teachers (pass/fail). Before this “Final practical”, they have already been several weeks in school during their teacher education process. Assessing the graduate teachers in authentic, real life settings gives, naturally, much more realistic a picture of the capabilities of managing the classroom, characteristics of the young teacher, as well as substance knowledge and pedagogical skills (see the discussion about the dimensions of the good teacher above in Section 5.2.1).

Psychometric Characteristics of the Tests

110. INICIA employed a number of volunteers, students and teachers to pilot the items across two phases during the test development phase⁷⁵. As can be seen from the reference these documents, in some cases no males were included in these pilot groups (for example, Evaluación1, pg.13); many items were omitted because of low item discrimination (for example, Evaluación4, 2013, 73–74). The INICIA provides a good reason why rigorous piloting of items is critical. The original tests, developed as observed earlier through a rigorous process would be good items for assessments. Since there were a limited number of participants in the pre-tests⁷⁶ and the limitations in including in the piloting actual student teachers, compels the need for checking the item parameters with the real dataset and this should have been done prior to summing up the final scores. Furthermore, it is to be recognized that reliability, for example, is not a stable characteristic of an assessment, but should always be estimated using the current set of test takers.

111. In what follows, the items are analyzed by using two approaches. These include the IRT modeling and the classical test theory. The classical analysis is bound to the tests and versions; the parameters are not strictly comparable over these variations. When the same item has been used as the linking item between the versions *A* and *B*, the mean of item discrimination is used as a common parameter. In the distractor analysis, the versions are kept separate. Within the IRT modeling, the one-parametric modeling, that is, Rasch modeling is used because of the limited number of cases in the datasets. This means that the item discrimination is not estimated by IRT modeling but one needs to accept the classical parameters; here the item-total correlation (*Rit*) and item-rest correlation (*Rir*) are used.

112. The analysis is done in five flavors: first, the overview of the tests is given by showing the quality of the tests graphically and comments on the flagged items. Second, the distractor-wise analysis is performed for the flagged items. Third, some rough ideas are given of the Differential Item Functioning (DIF) analysis. Fourth, the item difficulty parameters are tabled by using IRT modeling. Finally, the scores of the different tests are compared after equating the scores by using the IRT modeling. It is good to keep in mind that for a stable (or even meaningful) item analysis, somewhat 200 cases should be analyzed. However, since in the INICIA, especially the *PCD* tests, the number of test takers or cases is very sparse ranging from as little as 43 to 80. This results in unstable estimates.

Overall Quality of the Tests and Items

113. The overall quality of the items is evaluated on the basis of the classical item parameters - item difficulty and item discriminating power. As noted earlier, the higher the item discrimination power of the individual items, the higher will be the reliability of the test. Hence, more emphasis is put into item

⁷⁵ Please refer to Evaluación1, pg.13, Evaluación2, pg.25, Evaluación3, pg.18, Evaluación4, pg.33.

⁷⁶ Between 42 and 292 subjects in the individual assessments

discrimination than difficulty. Figures 9-20 illustrate the profiles of the tests for the various subjects. For purposes of exposition, we use the first of these tests, the *PCE-INICÍA*, to provide a more detailed and thorough analysis of this approach.

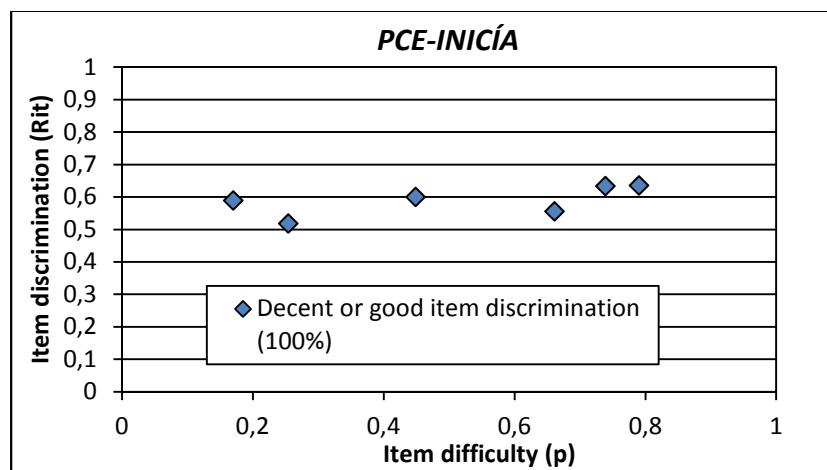


Figure 9: Relationship between item difficulty and item discrimination in the *PCE-INICÍA*

114. Figure 9 illustrates the relationship between item difficulty and item discrimination. Each square represents one item; in the *PCE-INICÍA*, there were six items: *Ortografía*, *Cohesión del texto*, *Vocabulario*, *Tesis*, *Estructura*, and *Argumentación*. The higher the square is located in the graph the higher is the item discrimination and hence, the more accurate the item. As a Pearson point-biserial correlation, the item-total correlation ranges from -1 to +1, where +1 is the perfect positive correlation and -1 the perfect negative correlation. In the case of *PCE-INICÍA*, all the items are exemplary from the point of view of discrimination. Even with the most difficult item (*Ortografía*), with the proportion of correct answers $p = 0.17$, the item discrimination is high, *Rit* is equal to 0.59 (refer to Table 12). This means that test takers who performed well overall in the assessment, also gained higher marks in the orthographic dimension of the assessment – even though their score was not high in this “sub-test” or item. The reliability of the *PCE-INICÍA* is low with an $\alpha = 0.64$, and the obvious reason for this is the brevity or shortness of the test ($k = 6$) and the reduced variance in the items. However, for a test of only six items, the reliability is decent.

Table 5 : Classical item parameters of *PCE-INICÍA*

Name/ Abbreviation	maximum value	item difficulty (p)	item discrimination (Rit)
Ortografía	6	0,170	0,589
Cohesión del texto	6	0,254	0,518
Vocabulario	6	0,449	0,600
Tesis	6	0,739	0,633
Estructura	6	0,790	0,635
Argumentación	6	0,661	0,555

115. We now turn to the other assessments. For reasons of clarity only the graphs corresponding to each assessment is presented below, and all the corresponding tables (as Table 6 above) can be seen in Appendix A.

116. The Básica tests (Figures 10 and 11) contain quite many low discriminating items (28% in PCD and 43% in PCP), however, no pathological ones. In both tests, even the highest values of item discrimination stay, in general, lower than $R_{it} = 0.40$. The length of the PCD-Básica ($k = 80$) causes the reliability to be quite high ($\rho = 0.81$ in version A and $\rho = 0.83$ in version B). In the PCP-Básica, almost half of the items are low-discriminating. Technically speaking, the combination of a short test ($k = 50$) and the low item discriminations ($R_{it} < 0.40$) causes the low reliability in PCP-Básica ($\rho = 0.68$ in both version A and B). Omitting/rewriting a couple of low-discriminating items may raise the reliability.

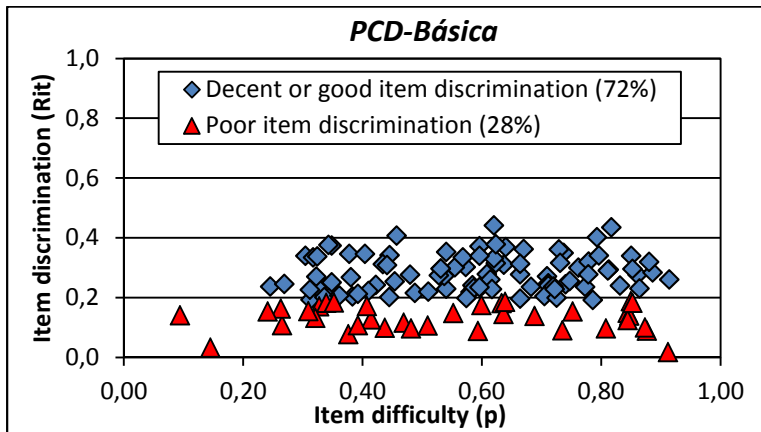


Figure 10: Item Discrimination and Difficulty of *PCD-Básica*

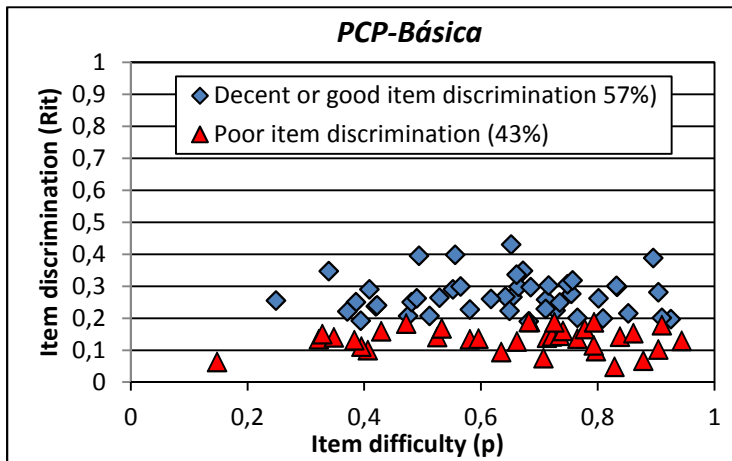


Figure 11: Item Discrimination and Difficulty of *PCD-Básica* and *PCP-Básica*

117. The PCD-Matemática and PCD-Física (Figures 12 and 13) are exceptions among the sub-tests of the INICÍA, and especially among the subject-wise tests. They contain few non-discriminating items (15% in the *PCD-Matemática* and 8% in the *PCD-Física*) and many highly discriminating items. The high item discriminations (up to $Rit = 0.68$ in the *PCD-Física* and up to $Rit = 0.56$ in the *PCD-Matemática*) and the fact that there are very few poor items causes the reliability to be high ($\alpha = 0,91$ in the *PCD-Física* and $\alpha = 0.88$ in the *PCD-Matemática*). From the item difficulty viewpoint, the tests cover the whole range of ability levels – and hence, presumably the tests could discriminate students at all difficulty levels. With some minor modifications, these two assessments could be improved even further, for example, by omitting or rewriting the low-discriminating items, and reliabilities could be raised from 0.91 to 0.92 and 0.88 to 0.89.

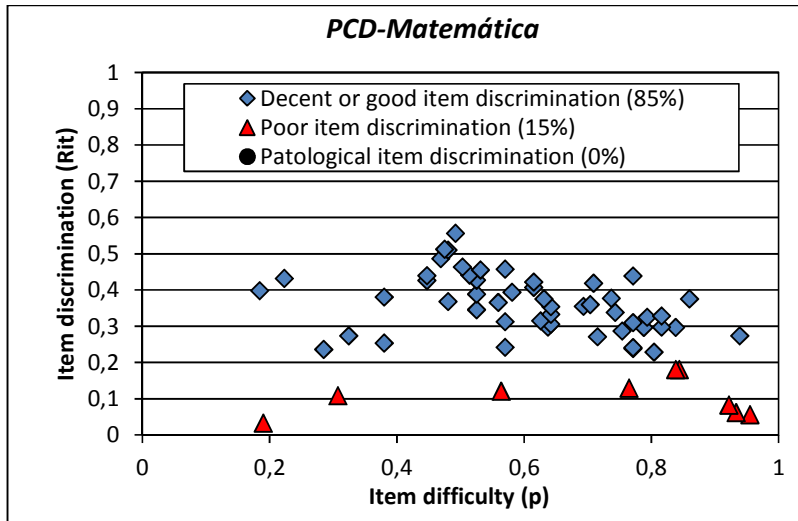


Figure 11 : Item discrimination and -difficulty of *PCD-Matemática*

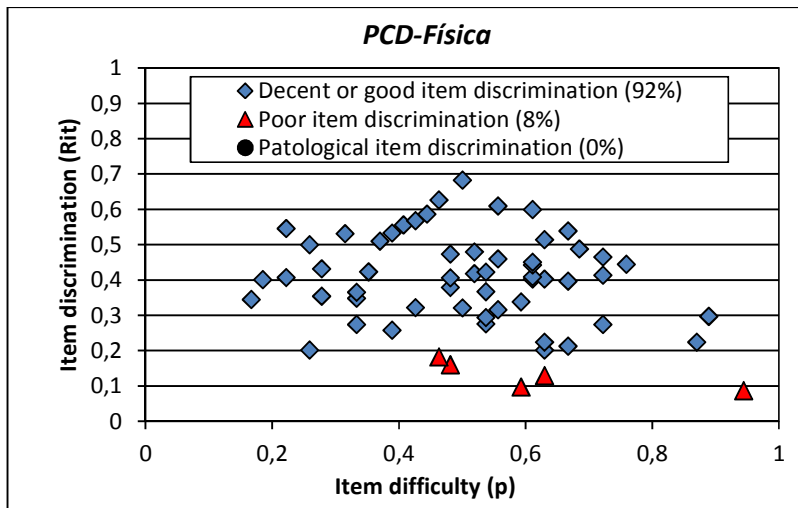


Figure 12 : Item discrimination and -difficulty of *PCD-Física*

118. Figures 14-17 below illustrate the relationship between item difficulty and discrimination for *PCD-Biológica*, *PCD-Química*, *PCD-Historia*, and *PCD-Lenguaje*. Compared with the *PCD-Matemática* and *PCD-Física*, these illustrate a greater number of poor or even pathological items. The black dots in the figures below represent pathological items and as can be seen, *PCD Biológica*, *-Química*, *-Historia*, and -

Lenguaje include several pathological items (3-10%). These items should have been detected and omitted before summing up the scores. The reliabilities are moderate ($\alpha=0.77$ in PCD Biologica, $\alpha=0.80$ in PCD-Química, $\alpha=0.75$ in PCD-Historia version A and $\alpha=0.72$ in version B, and $\alpha=0.71$ in PCD-Lenguaje version A and $\alpha=0.72$ in version B). The reliabilities would be raised by 0.02-0.03 units (0.77 to 0.80 and 0.80 to 0.82) just by omitting/rewriting the pathological items. Given the limited difficulty levels of the items in *PCD-Historia* and *PCD-Lenguaje* ($p > 0.30$), presumably the test is unable to discriminate high-achieving students from others very well - or at least the best students cannot show how good they would have been.

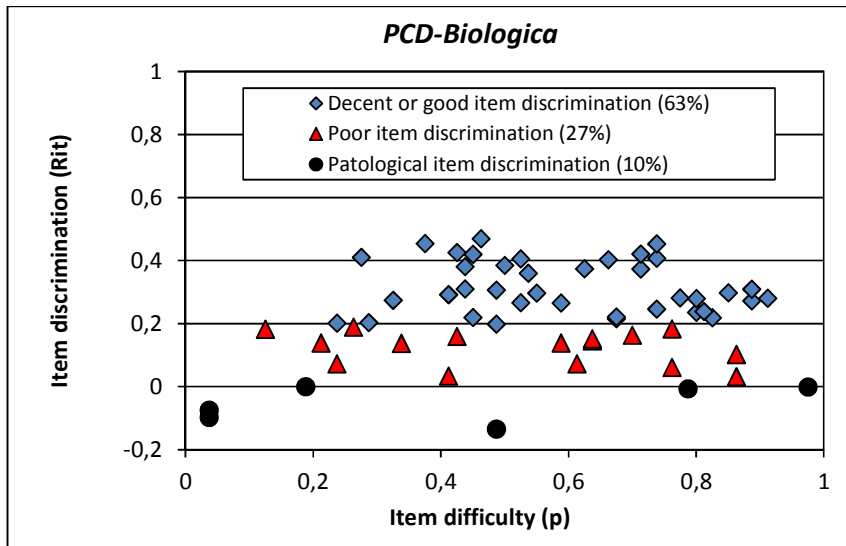


Figure 13 : Item discrimination and -difficulty of *PCD-Biologica*

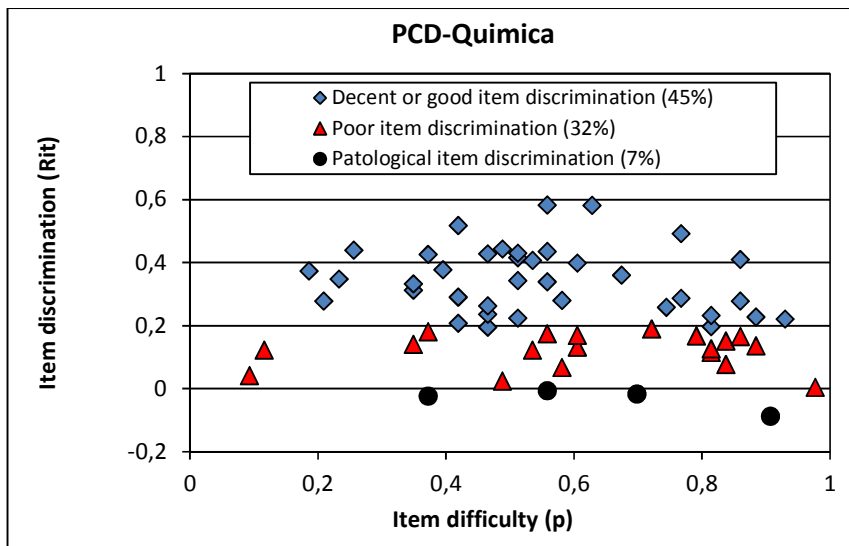


Figure 14 : Item discrimination and -difficulty of *PCD-Química*

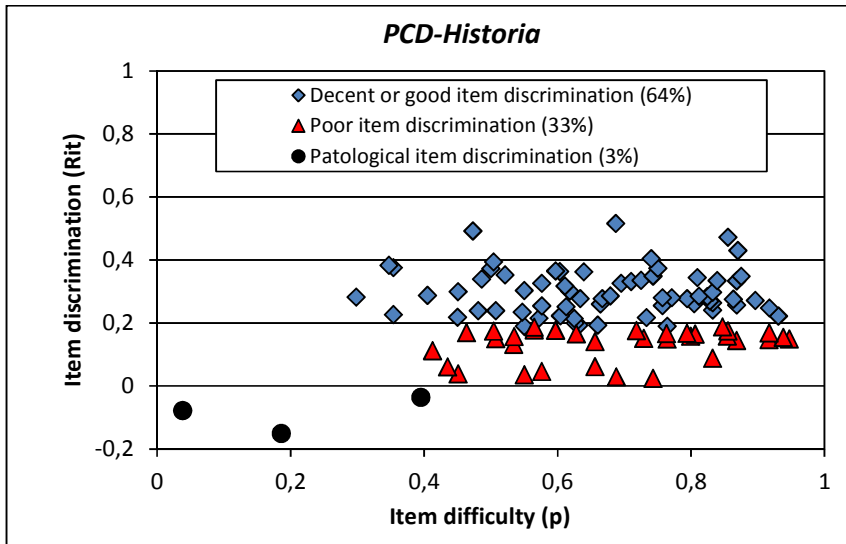


Figure 15 : Item discrimination and -difficulty of *PCD-Historia*

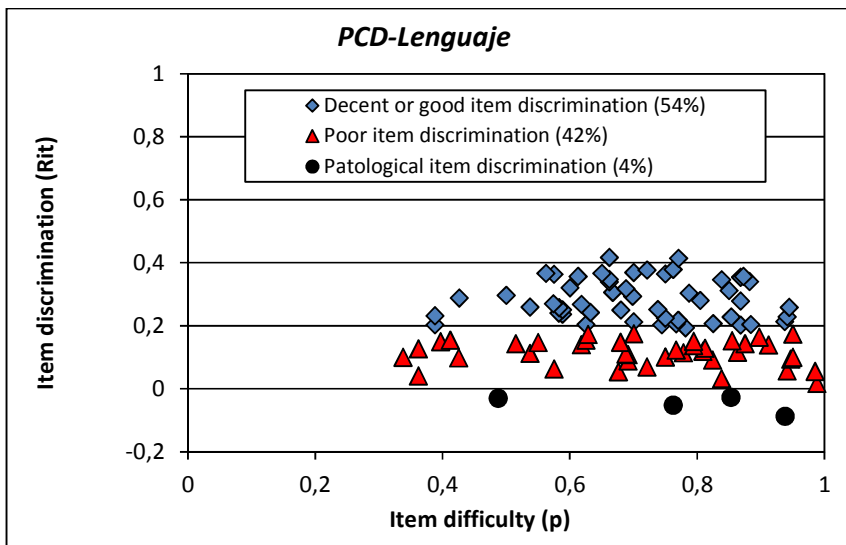


Figure 16 : Item discrimination and -difficulty of *PCD-Lenguaje*

119. The *Parvuria* tests include few pathological items (1–2%) and many low-discriminating items (41% in *PCD* and 34% in *PCP*). The reliabilities are quite low ($\alpha = 0.68$ in *PCD* version A and $\alpha = 0.71$ in version B and $\alpha = 0.68$ in *PCP* version A and $\alpha = 0.69$ in version B). Just by omitting/rewriting the pathological items, the reliabilities could be raised by about 0.02 units, or from 0.68 to 0.70 and if omitting/rewriting a couple of lowest-discriminating items, they would rise from 0.68 to 0.72 or 0.73. The overall level of item discriminating power is quite low ($Rit < 0.42$ in *PCD* and $Rit < 0.47$) and this evidently lowers the reliability of the test. Please refer to Figures 18 and 19 below.

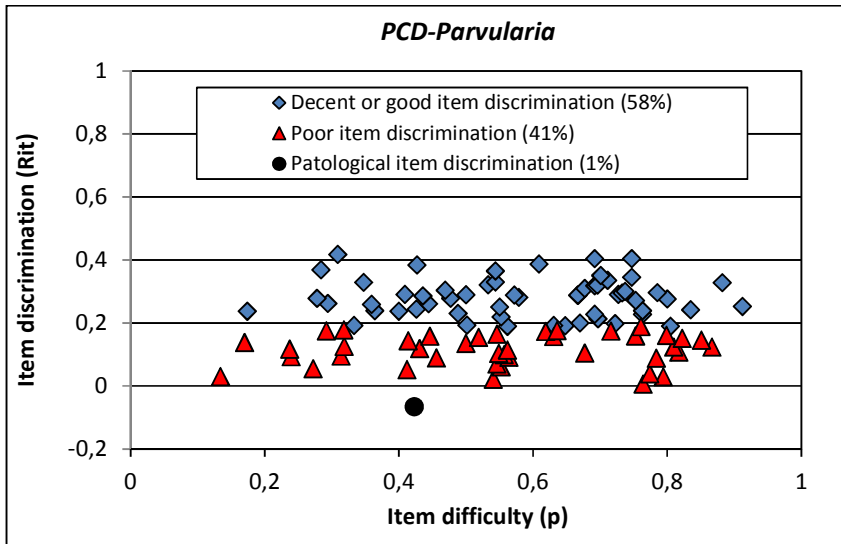


Figure 17 : Item discrimination and -difficulty of *PCD-Parvularia*

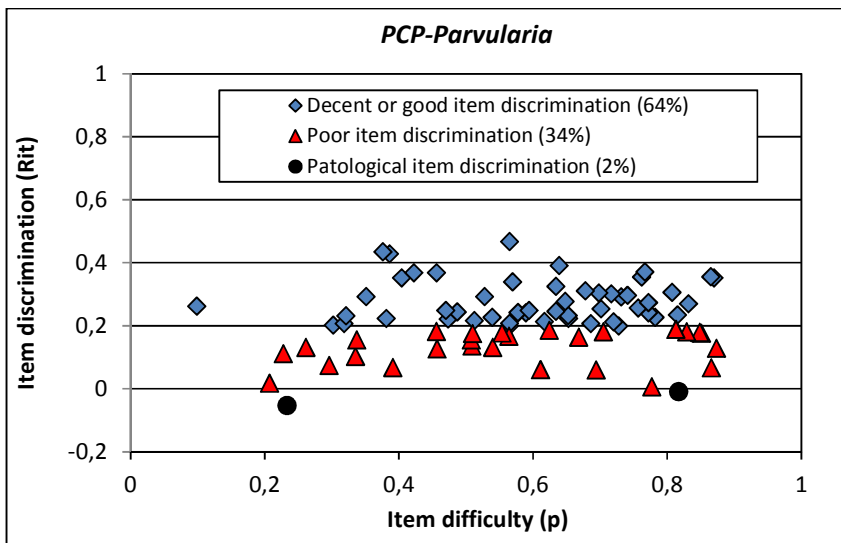


Figure 18 : Item discrimination and -difficulty of *PCP-Parvularia*

120. The *PCP-Media* test includes many low-discriminating items (32%) but has no pathological items. The reliability of the version A is quite low ($\alpha = 0.69$) and in version B it is decent ($\alpha = 0.77$). An obvious reason for the discrepancy is that, out of 28 low-discriminating items, 75% came from the version A. By omitting/rewriting a couple of lowest-discriminating items, $\alpha = 0.69$ could be increased to $\alpha = 0.70$. The overall level of item discriminating power is quite low, $Rit < 0.39$, except two items with somewhat higher value. This evidently lowers the reliabilities of the test. Please refer to Figure 20 below.

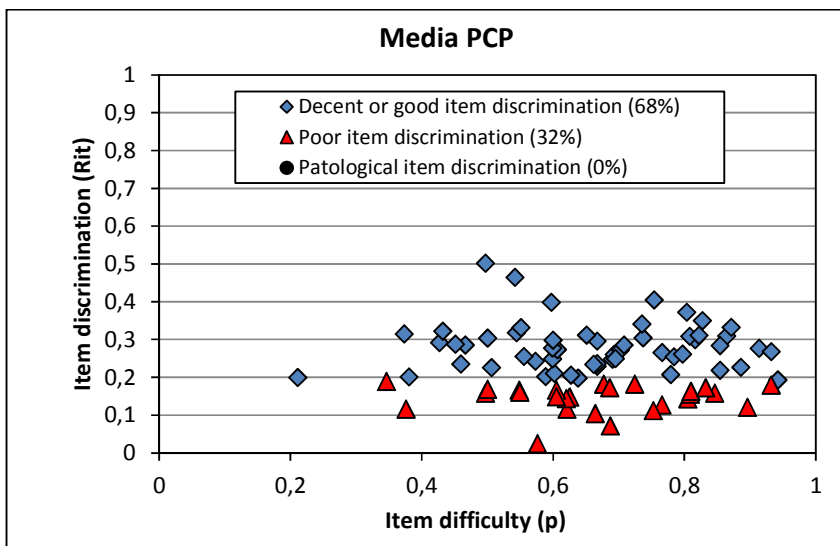


Figure 9 : Item discrimination and -difficulty of *PCP-Media*

121. After analysing all 915 individual items of the *INICÍA* test set, it is obvious that the assessment includes many low-discriminating items or *poor* items. There are 19 pathological items or about 2.1% of all items are pathological with negative item-total correlation. About 294 items or about a third of all items should have been omitted at the final phase because of very low item discrimination ($Rit < 0.20$). For future tests, it is recommended to omit or rewrite these poor performing items to raise the standard of the tests or select new items instead of the poor and pathological ones.

Distractor Analysis Of The Items

122. The previous analysis shows that there are several low-discriminating items in the test sets. If omitting these items would radically lower the validity of the tests it would be best to rewrite or amend the items, rather than merely omitting them. In order to be able to do this, the distractor analysis was done using TIAPLUS software. The specifics concerning the items are voluminous in nature and are collected in Appendix 2. However, in order to draw conclusions on the items on the basis of the distractor-wise graphs, an example of a *good item* is introduced as a reference (please refer to Figure 21).

123. The figure shows an exemplary multiple choice (MC) item from *PCD-Química*. The item-total correlation is high ($Rit = 0.50$). The legend on the right hand side shows that the alternative C is the correct answer. The test-takers have been divided into four groups (1 to 4) on the basis of their achievement level. Each curve tells what proportion of test-takers at each achievement level selected a specific alternative. When the item is a discriminating one, the curve related to the key should be (more or less) monotonously increasing as here: the lowest level test-takers do not select this alternative but the best ones do select the right one. For this particular item, the lowest level test-takers seems to be distracted by another option, alternative B. One could summarize that a well-discriminating or well-behaving item is characterized by the following characteristics:

- i. the highest-achieving students should select the correct alternative more probable than the lowest-achieving students,
- ii. the highest-achieving students should not be distracted to the incorrect alternative(s),
- iii. there should be at least one *real* alternative for the correct one which attracts the lower achieving test-takers.

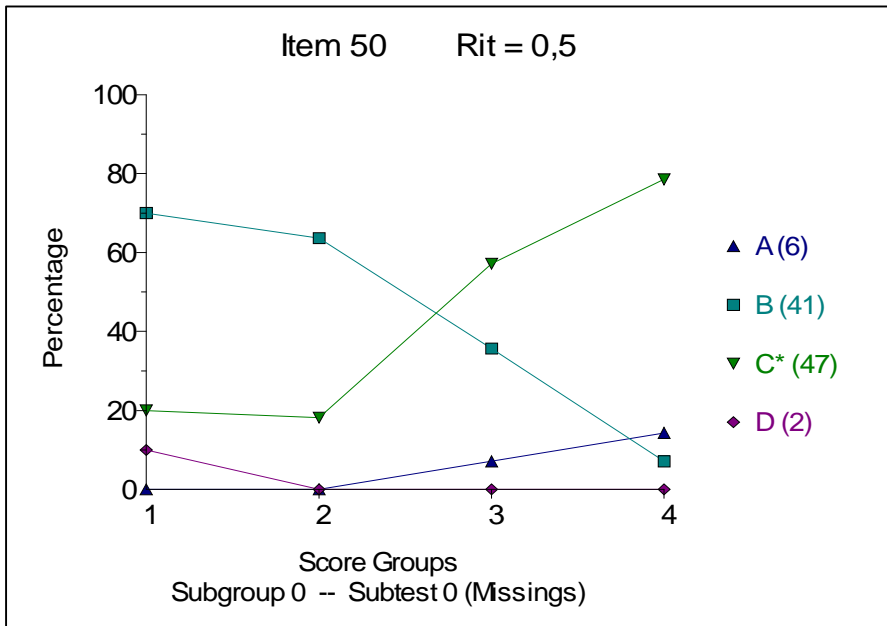


Figure 10 : An example of a graphical distractor-wise analysis of a good item

124. Mathematically, there are four indicators for a suspicious item and these are illustrated in what follows:

- i. the item total (or -rest) correlation (Rit or Rir) stays lower than 0.20
- ii. a distracter correlates as high as or higher with the test's rest score than the correct alternative, that is, $Rar \geq Rir$
- iii. the correct alternative does not correlate or even correlates negatively with the test's rest score, that is, $Rir \leq 0$, and
- iv. a distracter - test score correlation is suspiciously high, that is, $Rar \geq 10$.

125. As an example of a poor or even pathological item, Figure 22 shows an item from the same *PCD-Química* as above which looks good but which is technically a pathological item. The item shown below is flagged on the basis of three out of the four mathematical indicators discussed above in the context of a suspicious item:

- v. the item-total correlation is very low, that is, $Rit = 0.09$,
- vi. a distracter correlates higher with the test's rest score ($Rar = 0.33$) than the correct alternative ($Rir = 0.05$), that is, $Rar > Rir$,
- vii. a distracter-test score correlation is suspiciously high, that is, $Rar \geq 10$.

126. The graphical evaluation shows that the distractor B is monotonously increasing across ability levels, though the key does not identify distractor B as the correct answer, instead identifying alternative A as the right answer. Given that B would probably be the more likely answer; it is best to recheck the key and determine whether it has been erroneously coded. If the key *is* correct, then we can conclude that the item is poor and misleading and should be omitted.

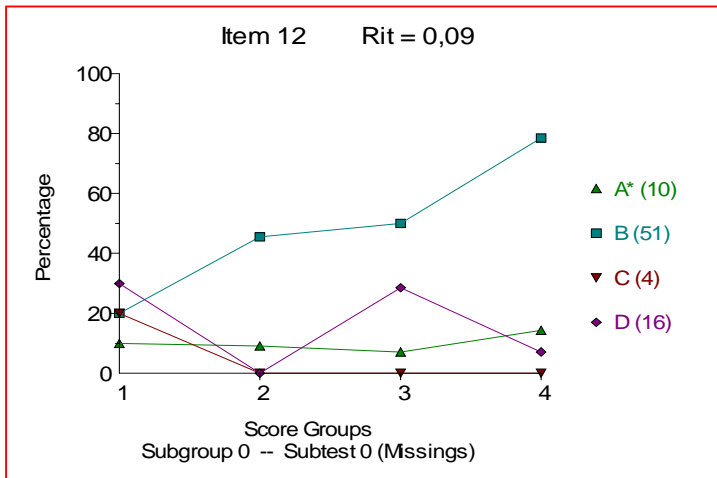


Figure 11 : An example of a distractor-wise analysis of a pathological item: a possible wrong key

127. Figures 23, 24 and 25 illustrate poor items of different kinds – found many times among the set of poor items identified. The item on Figure 23 is intuitively a very easy one to understand –*all the test-takers pick the correct alternative easily*. In numerous tests, items at the beginning of the test may be “motivating items” but in the middle of the test these seem to be too easy. The challenge with these kinds of items is that there is always one or several options which are not selected at all; these options are useless because even the weakest ones can easily out-select those. In the case of Figure 23, the weakest students tend to pick the correct answer but the better ones do not; making the item pathological.

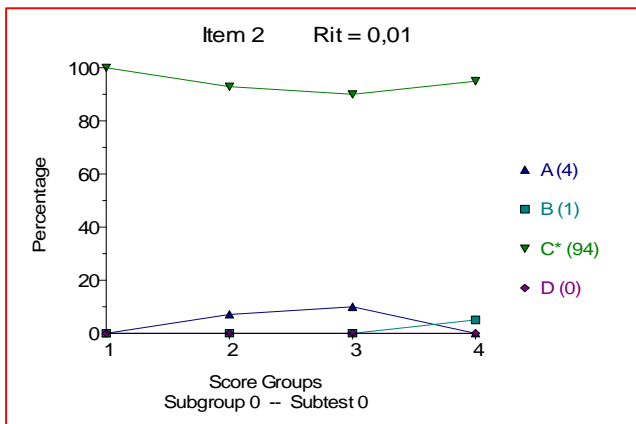


Figure 12 : An example of a distractor-wise analysis of a pathological item: an item with no alternative for the correct answer

128. A third type of pathological items involves an item flagged by all four indicators as shown in Figure 24. These four indicators illustrate that: (i) $Rit < 0$, (ii) $Rar > Rir$, (iii) $Rir < 0$, and (iv) $Rar > 0.10$. The findings show that that alternative D could be a good rival for the answer noted in the key - C. However, we conclude that here the point is that there is pathological guessing in the item. Thus we find that the weakest test-takers pick the correct alternative very easily (90% of them) but furthermore, the better students tend to select other options. These kinds of items need radical revisions or they should be omitted. Similar examples can be found especially, in PCD Química.

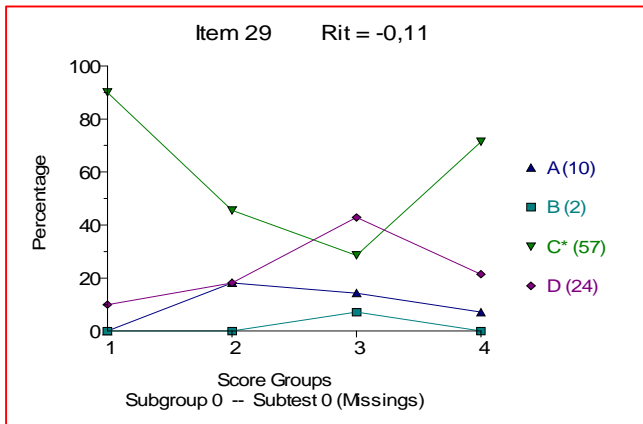


Figure 13 : An example of a distractor-wise analysis of a pathological item: Pathological Guessing

129. Still another kind of suspicious item type is the one which makes the best student confused regarding the right response. The item illustrated in Figure 25 is quite typical among the poor items: the best students are confused because there seems to be another (or several) correct answers. Here, potentially, the alternative C is a suspiciously good alternative for a correct answer. In the case, it is better to check the distractors; it is better to change the incorrect alternatives to be *more* incorrect, however, so that the weaker students would be attracted on this alternative. Naturally, if there really are *two* correct answers, the other one should be changed.

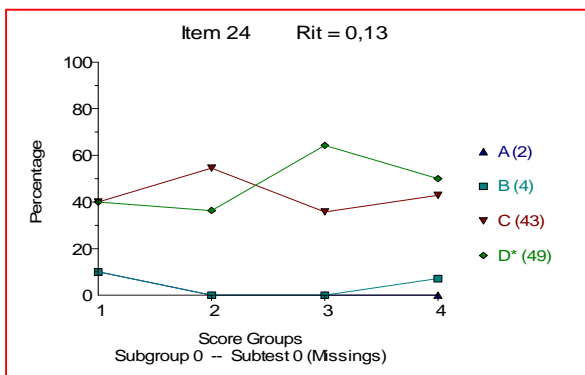


Figure 14 : An example of a distractor-wise analysis of a pathological item: Several Correct Answers

130. In what follows, only the flagged items from each test set are collected and commented. In many of the cases, the weakness in the item carries one (or more) characteristic(s) of the previous examples.

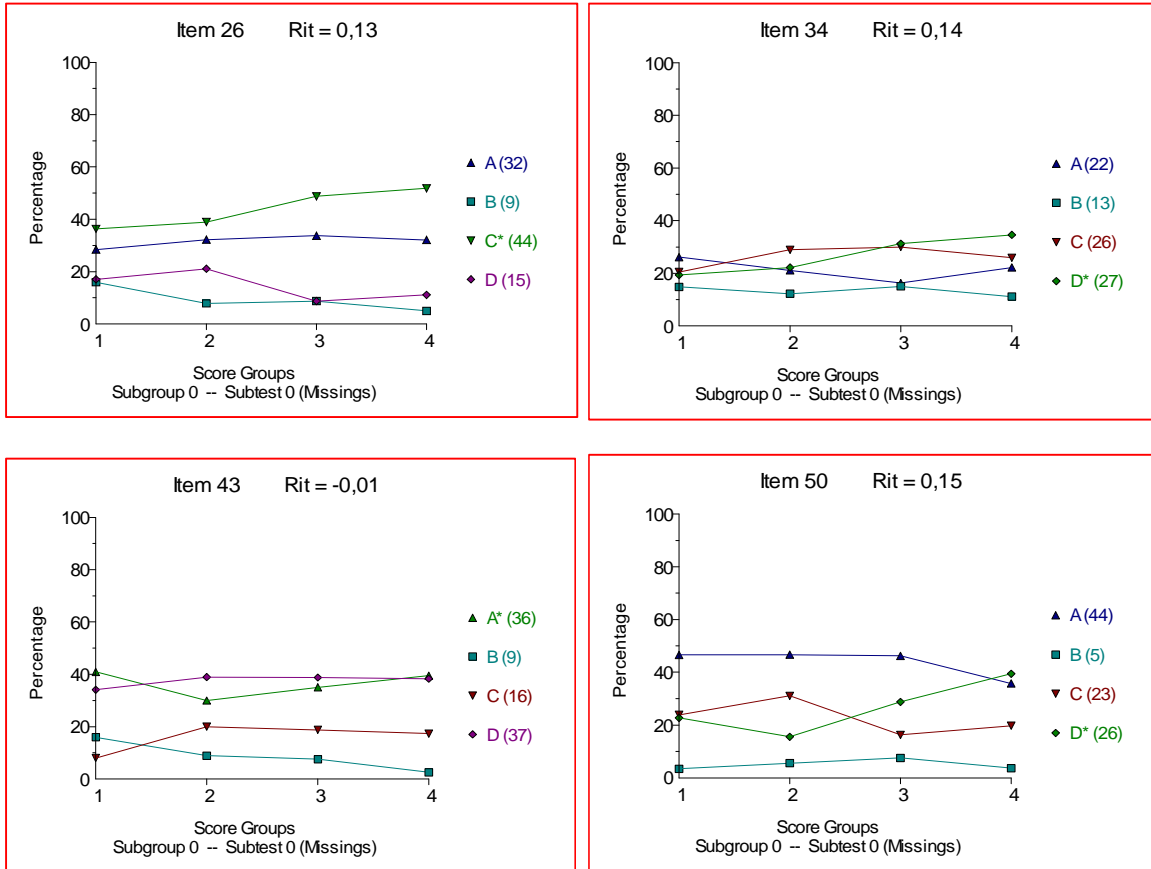
131. There are several characteristics in *PCD-Básica*⁷⁷ which emerge. Firstly, we note that for many items there are no options but the correct answer. For example, items 1, 2, 4, 12, 17, 45, 55, 56, 61, 64, and 76 (Version A) and items 1, 2, 10, 17, 61, 77, and 80 (Version B). In practice, this implies that there is at least one option which is not selected by any anyone and is typically not selected even by the weakest respondent. Such kinds of alternatives weaken entirely the item in question. However, by altering these distractors it is possible to improve the item. The second characteristic of the *PCD-Básica* which emerges – especially version B – is that the best students do not find the correct answer and are seen to be confused with another option. For example, items 26, 34, 43, 47, 50, 57, 62, 68, and 75 (Version A) and items 9, 13, 20, 34, 39, 43, 45, 53, 55, 56, 62, 64, and 75 (Version B). One would expect that for an item which is well

⁷⁷ See Tables B.1A and B.1B and the related graphs in Appendix B

developed, the best students would be able to identify the correct answers. So, when we find that the best students in fact select a distractor other than the correct option, it is better to check whether this alternative really *is* the correct one or there are multiple correct answers.

132. Among the items, none of them are seen to be pathological (except potentially 64 in the version A with potentially a wrong key) and only a few items with high guessing (potentially 43 and 50 in the version A and 1, 17, and 75 in the version B).

Version A:



Version B

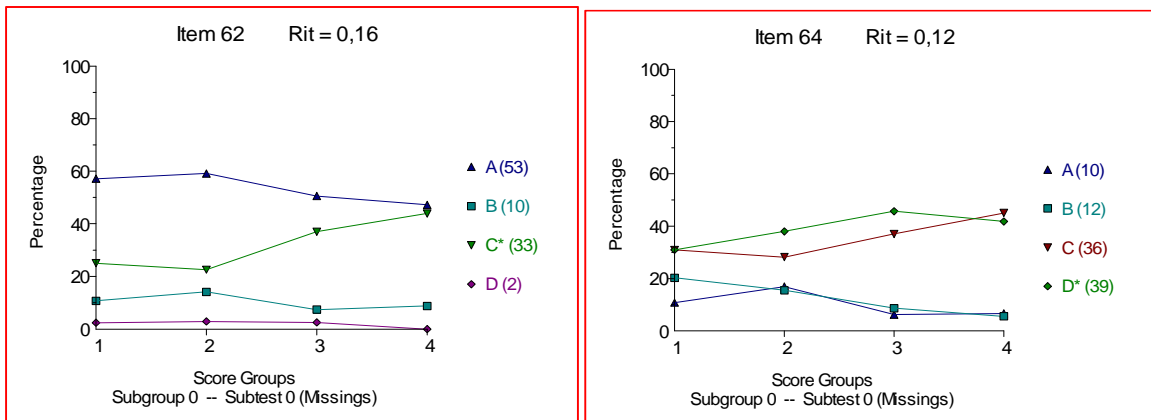
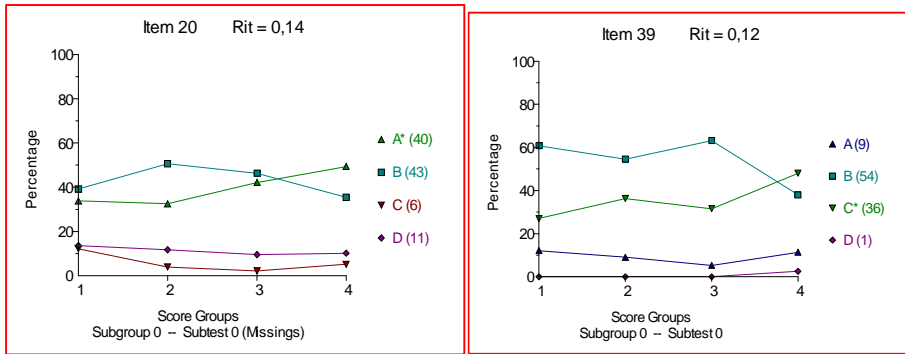


Figure 15 : A selection of suspicious or pathological items in *PCD-Básica*

133. In *PCP-Básica*⁷⁸ the key concern is that the weakest student group finds the correct answer too easily. Please refer to items 3, 5, 12, 16, 17, 34, 41, and 47 (Version A) and items 17, 19, 20, 23, 31, 42, and 48 (Version B). This means, in practice, that usually there is not an alternative, *really* distracting option for those who are not really knowledgeable of the contents of *PCP-Básica*. In many cases, this also means that there are at least one option which is not selected any anyone; it can be out-selected even by the weakest students. These kinds of alternatives are useless and just by altering these distracters may change the alternative better. Another characteristic of the *PCP-Básica* – especially version B – is that there seems to be two or more correct answers (items 20, 39, and 49 in the version A and 3, 7, 21, 23, 31, 33, and 50 in the version B). In the case, the best students do not find the correct answer; they are confused with (an)other option(s). The basic law is that the best students know the correct answer; when the best students select other than the correct option, it is better to check whether this alternative really *is* the correct one (or another correct one on the top of the real one). There are no items which are considered to be pathological ones. This holds except for item 3 in sersion A with potentially a wrong key and no items with high guessing.

Version A



Version B

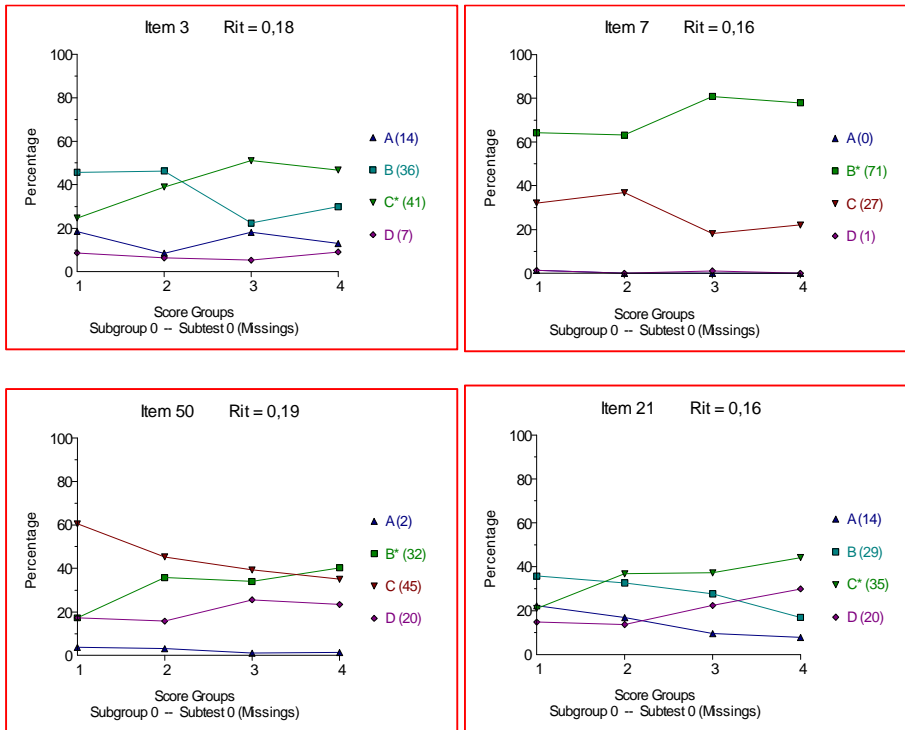


Figure 16 : A selection of suspicious or pathological items in *PCP-Básica*

⁷⁸ See Tables 5.3.2.2a and 5.3.2.2b and the related graphs in Appendix B

134. In *PCD-Biología*⁷⁹ the main concern is that the correct answer seems to be the only option. Please refer to items 3, 11, 35, 36, 44, 45, and 47. This means, in some cases, that there is no alternative, or *really* distracting option for those who are not knowledgeable of the contents of *PCD-Biología*. Therefore, the weakest student group finds the correct answer too easily. Such alternatives are not useful and by simply altering or modifying the distracters, it may be possible to improve the quality of the item. In *PCD-Biología*, there are also items where *two* (or several) options may be considered as the correct answer. We find that some of the best students are attracted to several of these options and consider them to be the right answers. Please refer to items 6, 10, 20, 33, 34, 40, 50, 52 and 53. Once again, the key assumption here is that the best students know the correct answer; and when systematically some of the best students select options other than the correct option it is better to check whether this alternative *really might be* the correct one (or the development of the items incorrectly permitted more than one correct option other than the identified one). Four of the items (33, 40, 47, and 52) seem to be pathological (see Figure 28); the analysis suggests that the key might not be correct, and hence it is best to review the key first; if the key is correct, then the items would need to be radically revised.

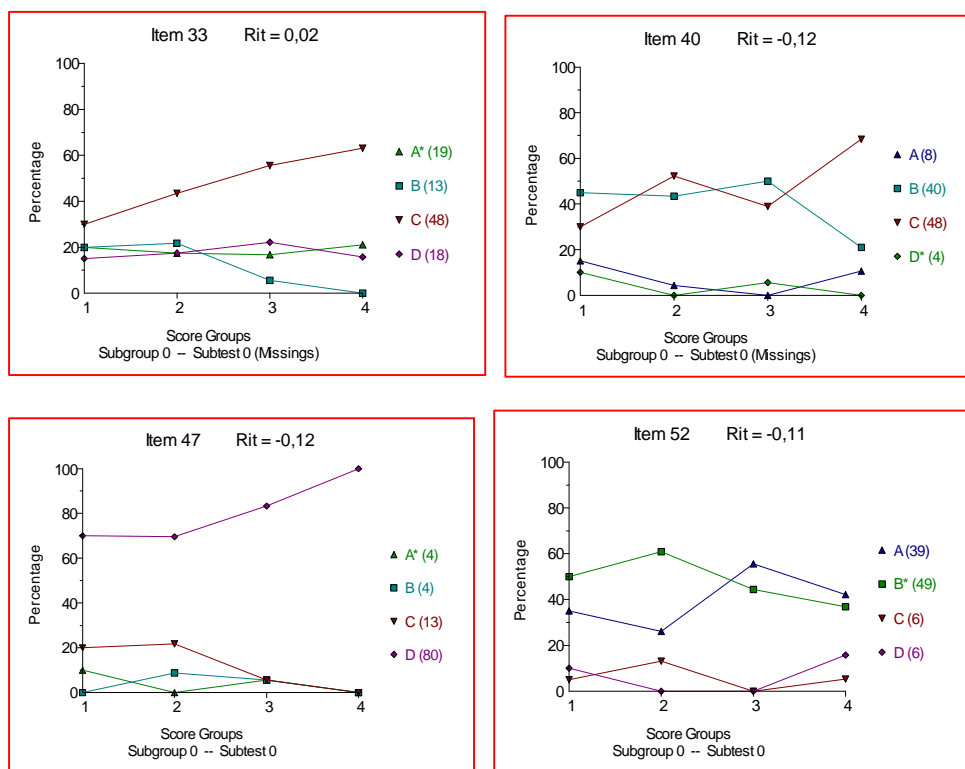


Figure 17 : A selection of suspicious or pathological items in *PCD-Biología*

135. In *PCP-Física*⁸⁰ there are only a couple of items with low item discrimination (for example 43, 44, 46, 56, and 59). Several items have a suspiciously high distractor-rest correlation though, that is, a distractor other than the correct alternative behaves as it would be the correct alternative. In most cases, this does not lead to low item-total correlation. In two cases, the low *Rit* in the items is caused by the fact that there are no options to the correct answer (items 43 and 56), which means that there is not an alternative, which *really* is a distracting option for those who are not really knowledgeable of the contents of *PCP-Física*. Furthermore, there is at least one option which is not selected by anyone and is typically out-selected even by the weakest students. These kinds of alternatives are not useful and by altering the distracters the item could improve its characteristics. In three items (44, 46, and 59) the best students are distracted by another option than the preset key; there seems to be two or more options for the correct answer – some of the highest ability

⁷⁹ Please refer to Table 5.3.2.3 and the related graphs in Appendix B

⁸⁰ Please refer to Table 5.3.2.4 and the related graphs in Appendix B.

students are attracted to these. In such cases, it is always best to check the key first; if the key proves to be correct, the items would need to be revised. None of the items show pathologically high guessing.

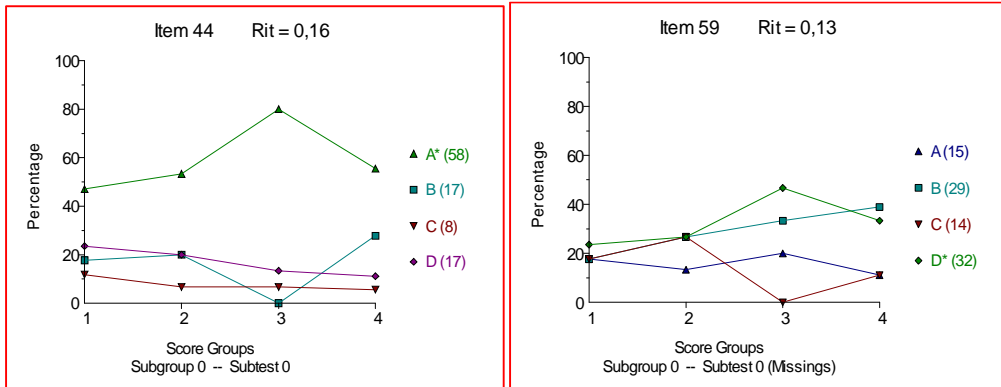


Figure 18 : A selection of suspicious or pathological items in *PCD-Física*

136. In *PCP-Matemática*⁸¹ there are only a couple of items with low item discrimination – quite many items of these (1, 3, 32, and 50) have low item discrimination because there are no options for the correct answer. In all cases, this means that there is at least one option which is selected by no one, that is, even the lowest level test-takers can out-selected these options. These kinds of alternatives are useless and just by altering these distracters may change the alternative better. In a couple of items (37, 47, and 50), there seems to be two options for the correct answer; the best ones are confused. It is better to check the key first; if the key was correct, the items need a radical revision. None of the items showed pathological guessing.

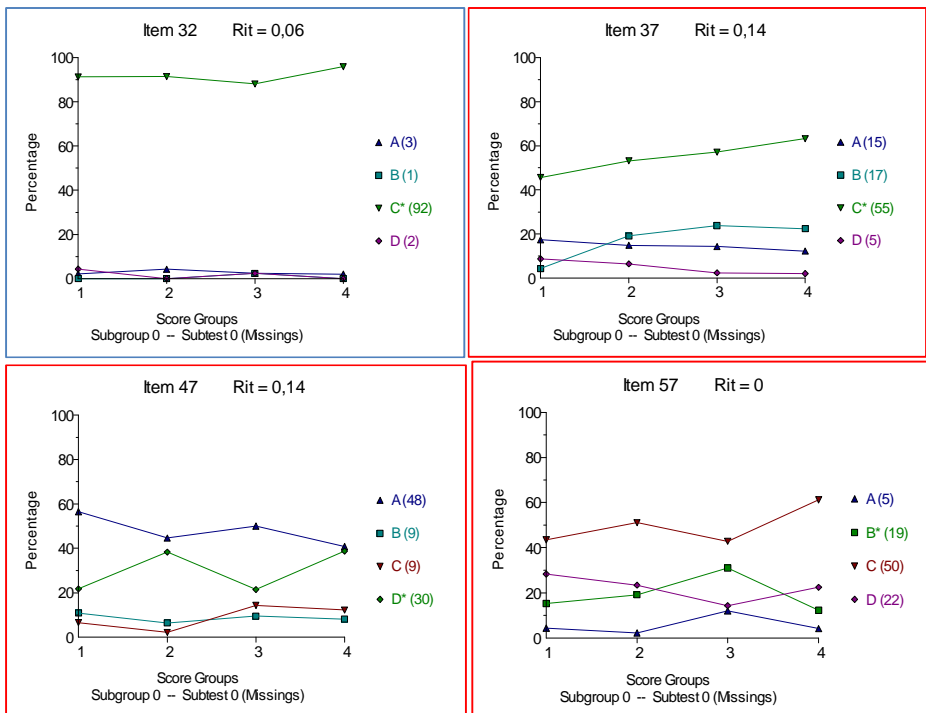
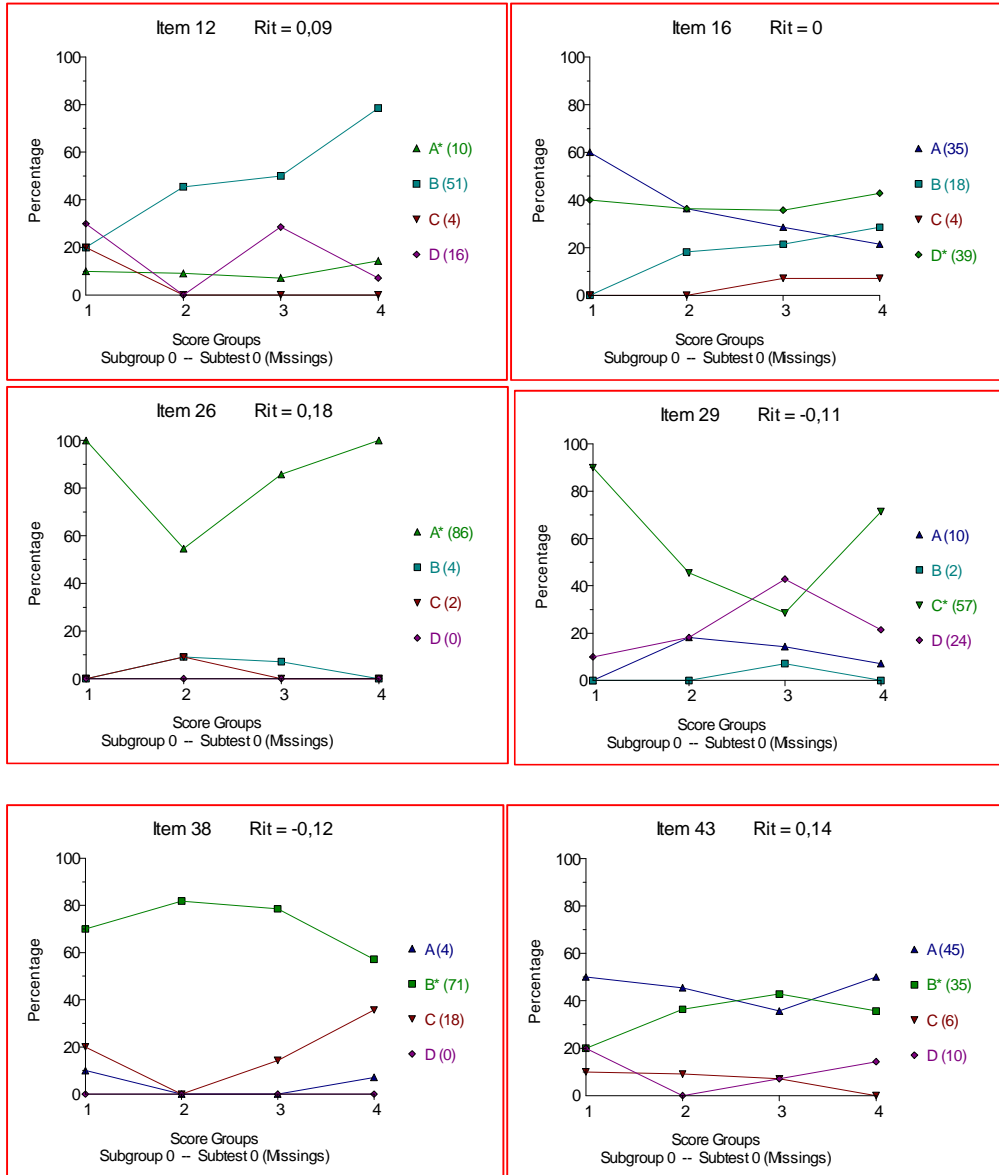


Figure 19 : A selection of suspicious or pathological items in *PCD-Matemática*

⁸¹ Please refer to [Table 5.3.2.5](#) and the related graphs in Appendix B.

137. In *PCP-Química*⁸², compared with *PCD-Física* and *PCD-Matemática*, there are quite many low-discriminating items and some really pathological ones. Some items (1, 2, 3, 22, 23, 26, 27, 35, 39, 46, 47, and 54) have low item discrimination because there are no options for the correct answer. In all cases, this means that there is at least one option which is selected by no one, that is, even the lowest level test-takers can out-selected these options. These kinds of alternatives are useless and just by altering these distracters may change the alternative better. Also, in quite many items (16, 17, 24, 29, 38, 43, 45, 51, and 57), there seems to be two or more options for the correct answer; the best ones are confused. It is better to check the key first; if the key was correct, the items need a (radical) revision. Two items (29 and 38) show a pathologically low item-total correlation and quite many of the items show a pathologically high guessing (1, 5, 17, 23, 26, 27, 29, 35, and 57). Refer to Figure 31.



⁸² Please refer to [Table 5.3.2.6](#) and the related graphs in Appendix B.

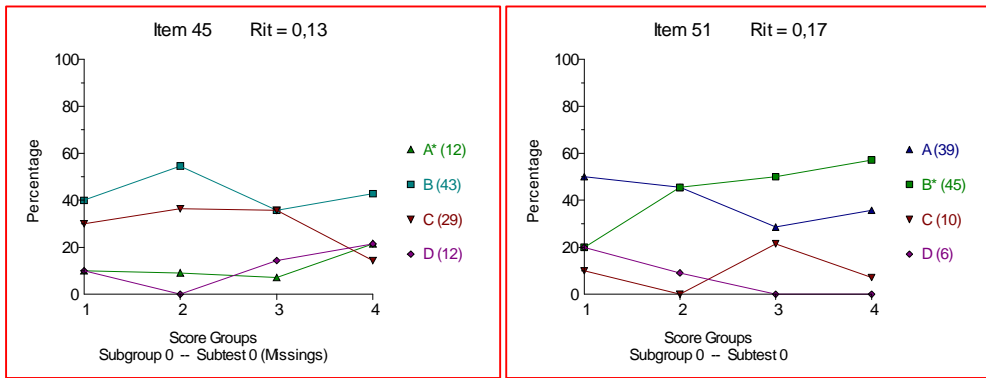
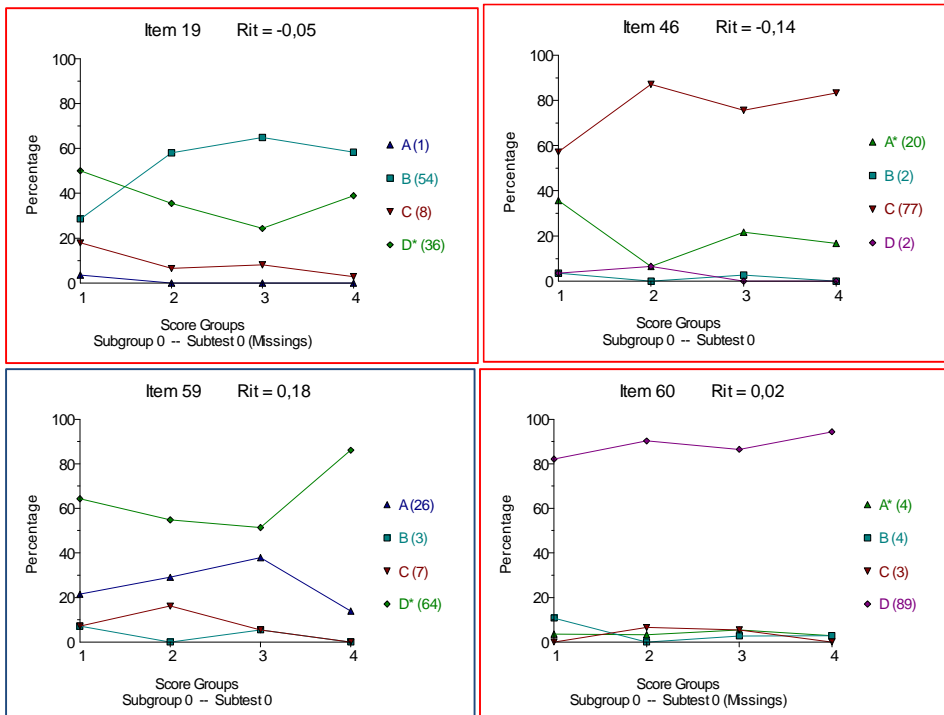


Figure 20 : Selection of suspicious or pathological items in *PCD-Química*

138. In *PCP-Historia*⁸³, compared with *PCD-Física* and *PCD-Matemática*, there are many more low-discriminating items and some pathological ones. For example, items 1, 10, 12, 36, 46, 50, and 60 (Version A) and items 10, 12, 21, 25, 27, 37, and 59 (Version B) have low item discrimination because there are no options other than the correct answer. In all cases, this means that there is at least one option which is selected by no one, that is, even the lowest ability test-takers will select out of these options. These kinds of distractors are not useful and the item can be improved by altering these distractors. In a significant number of items, for example, 6, 16, 17, 19, 23, 43, 49, and 51 (Version A) and items 16, 19, 33, 34, 37, 44, and 51 (Version B), there seems to be two or more options for the correct answer, leaving even the best student uncertain as to which is the right answer. Reviewing whether the key has identified the right answer might be a way of dealing with this issue and if this is found to be consistent, the item would need to be revised. Four items (29 and 38 in version A and 37 and 41 in version B) show a pathologically low item-total correlation and quite many of the items show a pathologically high guessing (Figures 32).

Version A:



⁸³ Please refer to [Tables 5.3.2.7a and 5.3.2.7b](#) and the related graphs in Appendix B.

Version B:

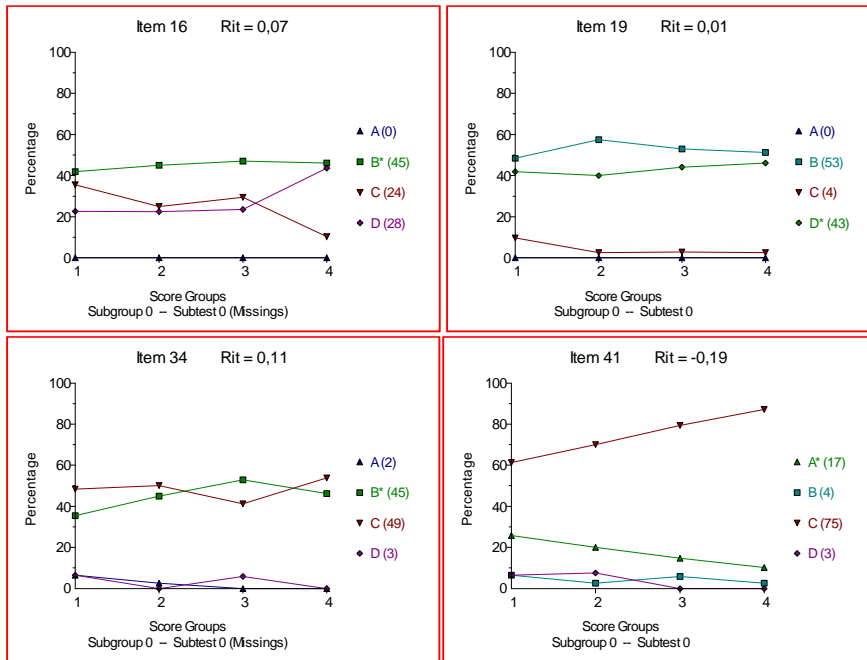
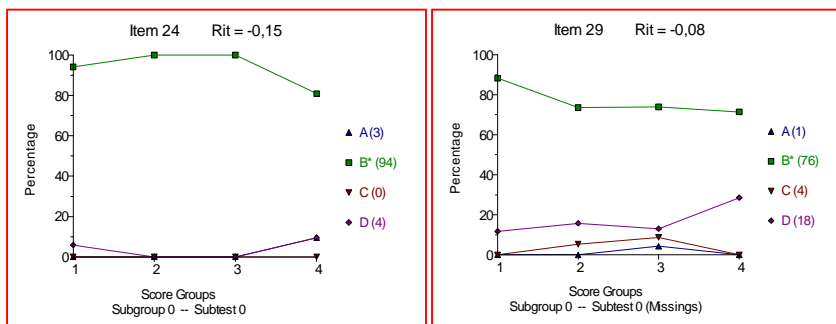


Figure 21 : Selection of suspicious or pathological items in *PCD-Historia*

139. Compared with *PCD-Física* and *PCD-Matemática*, in *PCP-Lenguaje*⁸⁴ there are several low-discriminating items and some pathological ones. Items 1, 5, 8, 24, 29, 32, 33, 36, 39, 40, 42, 43, 46, 49, and 53 (Version A) and items 1,2, 11, 30, 31, 34, 35, 37, 38, 39, 40, 42, 45, 46, 51, 53, and 56 (Version B) have low item discrimination because there are no real alternatives for the correct answer, implying that almost all students across the ability distribution get it right and there is at least one option that even the lowest level test takers are able to select out from choosing. Furthermore, items 14, 16, 23, 28, 29, 32, 52, 57 and 59 (Version A) and items 6, 9, 19, 24, 27, 30, 32, and 50 (Version B), there seems to be two or more options for the correct answer, resulting even in a set of the highest ability students erroneously choosing the incorrect option. Four items including 24 and 29 (Version A) and 1 and 53 (Version B) show pathologically low item-total correlation and many of items demonstrate pathologically high levels of guessing, including items 1, 23, 29, 46, and 56 (Version A) and items 2, 40, 45, 46, 47, and 51 (Version B). Refer to Figure 33.

Version A



⁸⁴ Please refer to [Tables 5.3.2.8a and 5.3.2.8b](#) and the related graphs in Appendix B.

Version B

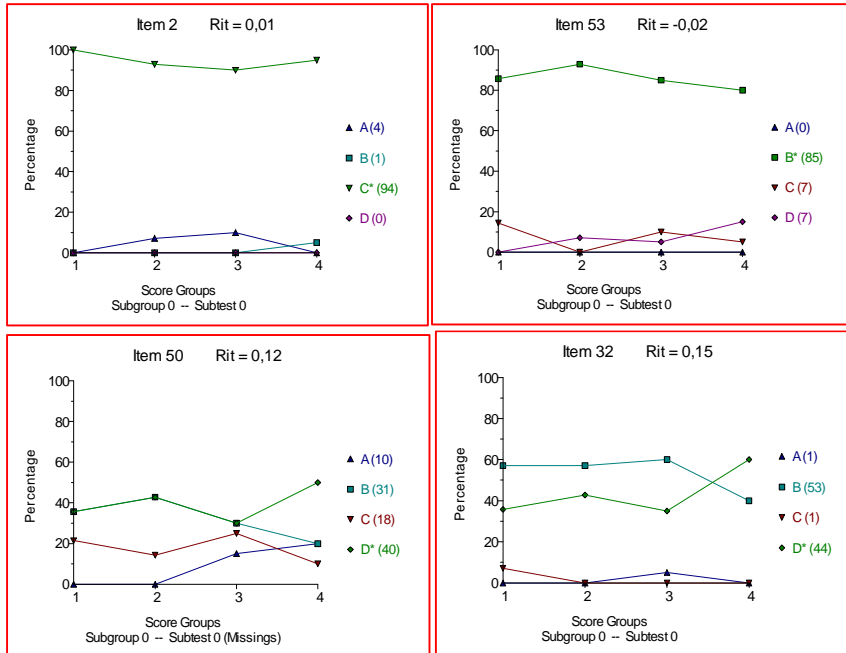


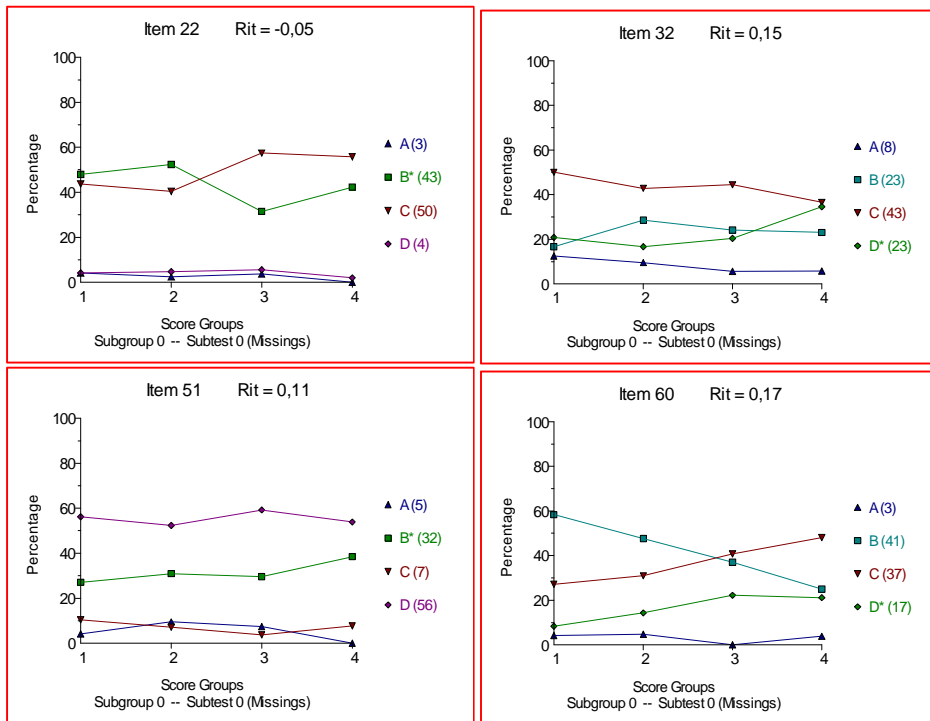
Figure 22 : Selection of suspicious or pathological items in *PCD-Lenguaje*

140. Compared with *PCD-Física* and *PCD-Matemática*, in *PCD-Parvularia*⁸⁵ there are several low-discriminating items and some pathological ones. Items 1, 2, 11, 36, 45, and 53 (Version A) and items 1, 2, 17, 24, 27, 28, and 36 (Version B) have low item discrimination because there are no alternatives options to the correct answer and the items are too easy for the test takers. Furthermore, many items including 8, 9, 22, 32, 35, 47, 51, 56, 57, and 60 (Version A) and items 10, 14, 16, 21, 35, 39, 51, 52, 56, 59, and 60 (Version B) seems to be two or more options for the correct answer.

141. One item in Version A (item 22) demonstrates pathologically low item-total correlation implying they are more likely to be answered correctly by less skilled test takers than by more skilled test takers, and items 45 and 53 (Version A) and items 24 and 52 in (Version B) show a pathologically high guessing.

⁸⁵ Please refer to Tables 5.3.2.9a and 5.3.2.9b and the related graphs in Appendix B.

Version A



Version B

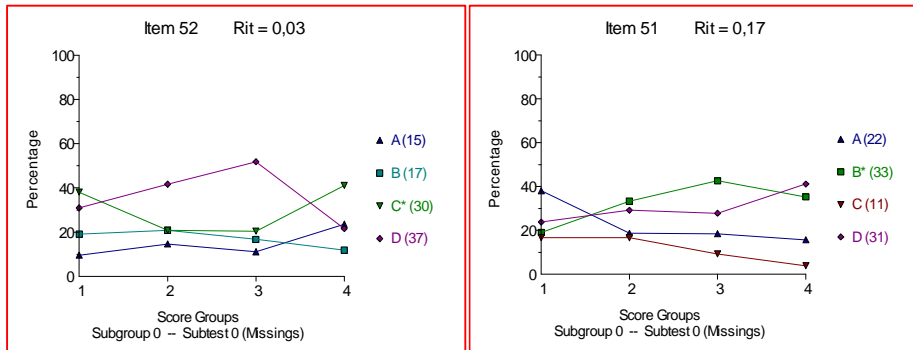


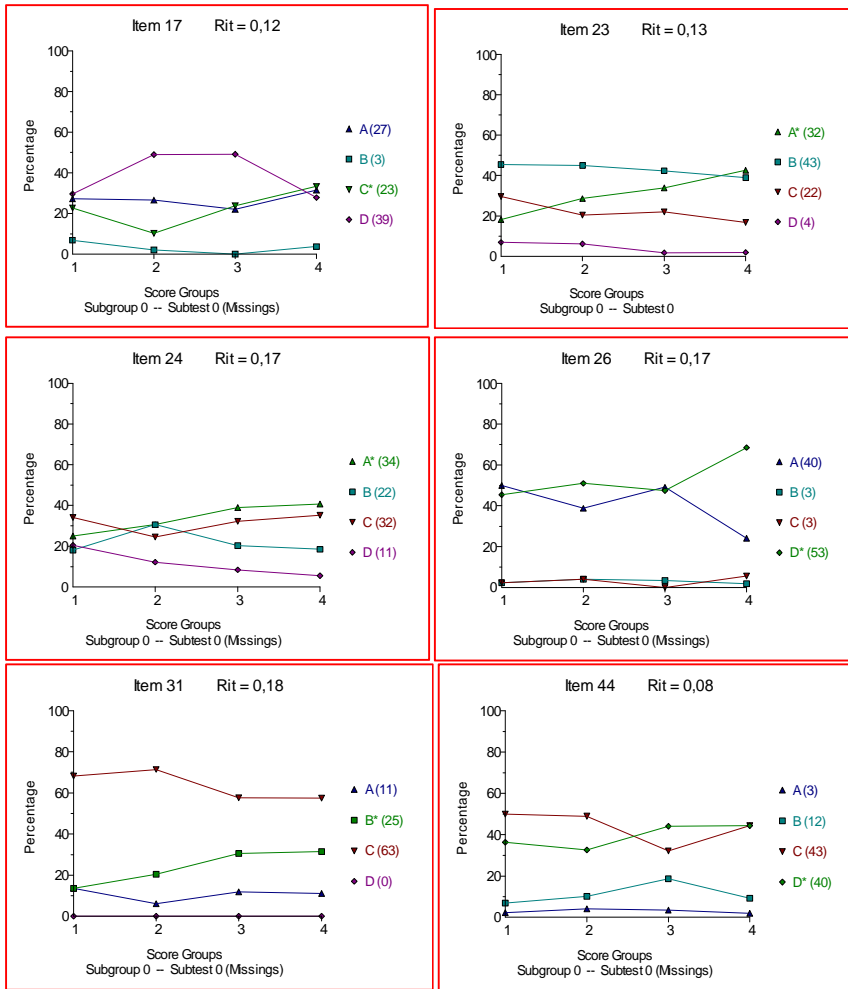
Figure 23 : Selection of suspicious or pathological items in *PCD-Parvularia*

142. Compared with *PCD-Parvularia*, in *PCP-Parvularia*⁸⁶ there are fewer number of low-discriminating items and only a few pathological ones. Items 12, 18, and 50 (Version A) and items 18 and 46 (Version B) have low item discrimination and these items are too easy for the population of test takers in that an overwhelming majority of the test takers are able to get the right answer. Furthermore, many items 17, 23, 24, 26, 31, 34, 42, 44, and 48 (Version A) and items 6, 17, 22, 23, 26, 30, 32, 35, and 48 (Version B), there seem to be two or more possible correct answers and this leads to even the more skilled students being confused as to the correct choice.

3. One item 17 (Version B) shows a pathologically low item-total correlation and two items 17 (Version A) and 17 (Version B) show a pathologically high guessing.

⁸⁶ Please refer to [Tables 5.3.2.10a and 5.3.2.10b](#) and the related graphs in Appendix B.

Version A



Version B

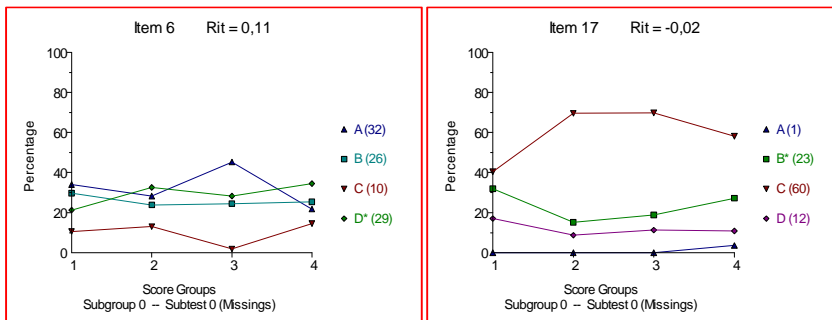


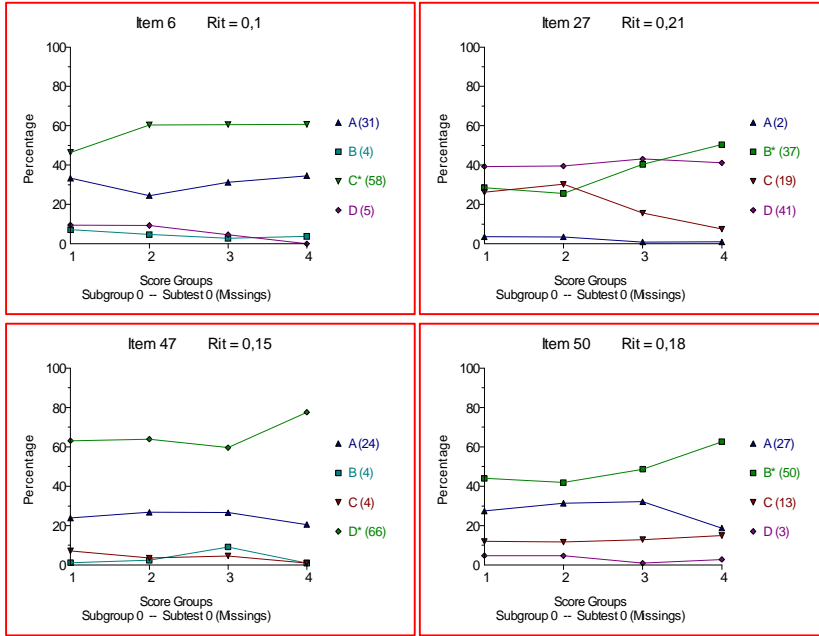
Figure 24 : Selection of suspicious or pathological items in *PCP-Parvularia*

143. Compared with *PCD-Parvularia*, in *PCP-Media*⁸⁷ there are fewer low-discriminating items and no pathological ones. Several items 8, 13, 25, 26, 28, 41, and 46 (Version A) have low item discrimination because the items are too easy for the test taking population and almost all students get the right answer

⁸⁷ Please refer to [Tables 5.3.2.11a and 5.3.2.11b](#) and the related graphs in Appendix B.

including the lowest skilled students. In several other items 6, 27, and 50 (Version A) and items 2, 20, 27, 37, and 47 in (Version B), there seems to be multiple options for the correct answer and even the more skilled test takers are confused as to the correct option. Two items 2 and 47 (Version B) show a pathologically high guessing.

Version A



Version B

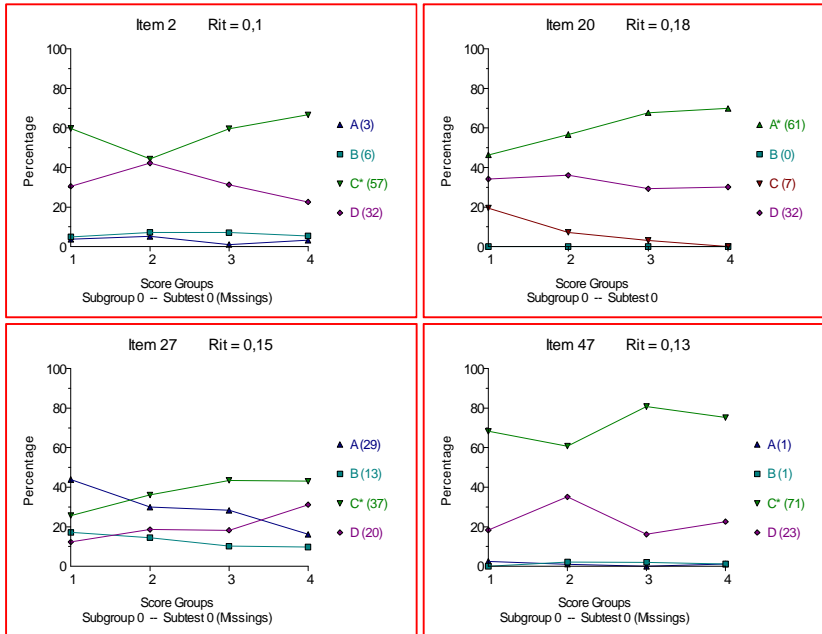


Figure 25 : A Selection of suspicious or pathological items in *PCP-Media*

Summary of Distractor Analysis

144. All in all, there are four kinds of challenges in the flagged items. In many cases, the items are too easy for the test taking population and there is really only one alternative to select – which happens to be the correct one. In these items, even the weakest students know, just recognize, or guess the correct answer too easily and, hence, the low item discrimination. In such cases we also observed that there are usually one or more alternatives never selected. It may be worthwhile to rewrite the items so that these alternatives are amended, if possible, to be more attractive so that the weakest students would select those and thereby strengthening item discrimination. Another commonly seen challenge is that there seems to be several correct answers which attract the best students. The ability to discriminate essentially entails that the strongest students should be more likely to arrive at the right answers while the weaker test takers should have a lesser chance of getting the right answer for each item across the item difficulty curve. In many items of the *INICÍA*, this does not happen. It may be worth considering revising (or at least checking) the items so that there really are not alternatives of a kind which can be (partly) correct and which the best test-takers pick because, they may be correct ones. Two less common challenges are connected by the fact that the weakest students seem to guess the correct answer too easily. In some cases, this evidently leads to the negative item-test correlation. Obviously, these items should be omitted or rewrite.

Differential Item Functioning (DIF) Analysis

145. The number of cases is, in most datasets, too sparse to perform a proper DIF analysis even for the smallest number of the comparable groups, that is, when comparing two groups. In the dataset, only one DIF analysis was done by using the Mantel-Haenszel (1959) statistics: The items were tested on the basis of the variable *Tipo de evaluado* which has two values: 1=*Egresado de pedagogía* and 2=*Beca Vocación de Profesor o Enseña Chile*.⁸⁸ In many cases, the number of cases in the group *Beca Vocación de Profesor o Enseña Chile* was sparse (less than 10% of the cases) which evidently affects the result. In most datasets, the classical thumb rule, note by Heuvelmans (1998, 5), of the ratio of 1:5, that is, five times more cases than items does not hold in the dataset. For each 60 item test, there should be round 300 cases in order to get sensible or stable results. Now, when the number of cases is 80 or less, it is good to be critical and careful with the results.

146. The original Mantel-Haenszel (MH) statistic is based on Chi Squared statistics; here it is converted to the Standard Normal distribution fractions. Statistically significant DIF would require values over 1.96. None of 915 items showed this high value. Hence, from the statistical viewpoint, none of the items show DIF. The graphical analysis, however, shows grave discrepancies between the groups. The MH statistics and the most suspecting DIF cases are collected in Appendix 3.

Item Parameters From The IRT Modelling

147. The difficulty levels of the items (B parameters in IRT modelling) are not necessarily interesting from an evaluation viewpoint. They are used, however, in equating the test scores (see following section). Item difficulties with the related standard errors are tabled in Appendix A. Item Characteristic Curves (ICCs) might have been informative from the guessing point of view. However, the distractor analysis in Section 4.3.2 and Appendix 2 tell the same information and hence, ICCs are not presented here. In the whole test set of 915 items, the item difficulties range from $B = -4.082$ to $B = 3.14$. The distribution of the item difficulties is geared toward easier items rather than difficult items (Table 13). This is not necessarily a problem.

⁸⁸ There were two other interesting variables to use in the DIF: *Año de egreso pedagogía categoría*, with categories 1=*Egresado 2010*, 2=*Egresado 2011*, 3=*Egresado 2012*, 4=*Beca Vocación de Profesor*, and 5=*Enseña Chile* and *Participación en Pilotaje* with categories *no piloto/piloto*. The former included too many categories and the latter applies barely one percent of the *PCE-INICÍA* participants.

However, from the test construction point of view it would have been better to allow the really good test takers the opportunity to show how good they are. Now it seems that the three most difficult items (see Bio_AD47, Bio_AD40, and His_AD40 in Appendix 2) are flagged as pathological ones; the item discrimination is negative and the percentage of correct answers is $p < 0.04$.

Table 6 : Distribution of B parameter values in *INICIA*

B	Description	Frequency	%
< -2.50	Very easy	19	2,1
-2.50 - -1.50	Easy	114	12,5
-1.51 - -0.50	Easy mediocre	290	31,7
-0.51 - +0.50	Mediocre	347	37,9
+0.51 - +1.50	Difficult mediocre	123	13,4
+1.51 - +2.50	Difficult	18	2,0
> +2.50	Very difficult	4	0,4

Equated Scores Over Tests

148. Maybe the most important question of all is whether the reporting categories (“insufficient”, “sufficient”, and “outstanding”) are fair for all test takers. In the most unfair case, the test-taker takes a test which is more difficult in comparison with the other tests, gets low score, and is labelled as “insufficient” – not because of being at the insufficient level but – because of a more difficult test or test version. Another student with the same achievement level, who took an easier test or test version, would be labelled as a “sufficient” one. In this Section, the test difficulties are evaluated on the basis of the equated scores.

149. The test equation is done on the basis of linking the tests with each other by the *PCE-INITIA* and by using the IRT modelling. It is essential that the item difficulties are first calibrated at the same scale (Section 4.3.4). After calibration, the latent, sample-free, ability level (Theta, θ) is estimated for each test score in each test and version. Theta tells how much achievement is needed for gaining each score. The raw scores are not comparable over the tests but the Theta values are. Hence, for example, the average Theta in the population ($\theta = 0.00$) can easily be compared over the tests and versions. The reference scores at three levels of achievement (Exceptionally low⁸⁹, Mediocre, and Exceptionally high⁹⁰) are collected in Table 13 (see more exhaustively in Appendix 4).

⁸⁹ The “Exceptionally low” is not something uniformly fixed. The boundary of 1.5 standard units below the average has been used as the boundary when assessing the exceptionally low-levelled students in the compulsory education in Finland (for example, in Räsänen & Närhi, 2013; Räsänen, Närhi & Aunio, 2010).

⁹⁰ Obviously, also the boundary for the “Exceptionally high” is not something uniformly fixed. The boundary of +1.50 standard units above the average is used here for the symmetrical reasons.

Table 7 : Reference scores of the components of the *INICÍA*

Set ¹	PCE ⁴	PCP			PCD							
Test ²	INICÍA	Bas	Med	Par	Bas	Par	Len	Mat ⁴	Bio ⁴	Qui ⁴	Fis ⁴	His
Numerus ³	1,824	669	754	295	663	289	80	179	80	43	54	131
Mean of θ												
Version A	0.08	0.12	0.19	-0.04	0.13	-0.06	0.37	0.21	0.09	0.14	0.09	0.00
Version B		0.05	0.16	-0.07	0.05	-0.05	0.58					0.15
Score at $\theta \leq -1.5$												
Version A	7	15	15	14	20	16	19	17	15	16	12	20
Version B		16	16	14	21	16	20					20
Score at $\theta = 0$												
Version A	18	32	32	30	44	36	39	36	33	34	31	40
Version B		33	33	30	47	35	40					39
Score at $\theta \geq +1.5$												
Version A	29	44	44	43	65	51	51	51	48	49	48	53
Version B		44	45	43	68	54	54					53
Maximum score	36	50	50	50	80	60	60	60	60	60	60	60

- 1) PCE = *Prueba de Comunicación Escrita*, PCP = *Prueba de Conocimientos Pedagógicos*, PCD = *Prueba de Conocimientos Disciplinarios*
- 2) Bas = *Basica*, Med = *Media*, Par = *Parvularia*, Len = *Lenguaje*, Mat = *Matematica*, Bio = *Biologia*, Qui = *Quimica*, Fis = *Fisica*, His = *Historia*
- 3) Combined version A + B
- 4) Only one version or parallel tests is in use

150. It is evident that the individual tests and test versions are not at the same difficulty levels (Table 14). The mediocre test-taker with $\theta = 0.00$ would gain in the *PCD-Física* only 31 points while, with the same latent ability level, the test-taker in the *PCD-Historia* and in *PCD-Lenguaje* 40 points even though the maximum values of the tests are the same. The same holds also at the boundary of exceptionally low-levelled test-takers ($\theta = -1.50$); a test-taker who would be, objectively taken, at the boundary of $\theta = -1.50$, would gain only 12 points in *PCD-Física* but 20 points in *PCD-Historia* and in *PCD-Lenguaje*. At the upper boundary of exceptionally high-levelled test-takers ($\theta = +1.50$), the differences between the test scores seem smaller (5–6 points) than in the lower level benchmarks (8–9 points). Because the scores differ from each other, it would have been profitable to equate the scores before calculating the reporting categories. This challenge is handled in the next section.

Adequacy And Comparability Of The Reporting Categories

151. The final judgments of the graduate teachers to be “insufficient”, “sufficient”, and “outstanding” are made on the basis of equated total score. The logic of the transformation from the original score to the equated score and to the standard deviation (z-score) in the datasets is not obvious though. In *PCE-INICÍA*, the logic differs from the other tests; the written thesis was categorized into grades of “*pass*” and “*fail*”.

152. Judging the graduate teachers on the basis of the norm-referenced test is a challenging task. Because there are no absolute criteria where to set the boundaries, they need to be negotiated. Even then one may ask relevant questions such as: *Who* decides where the boundaries are and on what basis? Shouldn't *all* the candidates know *all* the important things? *Who* decides what *is* important to know? In the norm-referenced testing, it may happen that *all* candidates are good enough in an absolute sense but the norm always points out some test-takers to be the lowest ones and the others to be the highest ones. Hence, the boundaries for “insufficient”, “sufficient”, and “outstanding” are not fixed in an absolute sense.

153. For the graduate teacher, the boundary of “insufficient” may be more crucial than being “outstanding”. The first thing that fixes ones attention on the final judging of the graduate teachers is the relatively high boundary for “insufficiency” or “failing”. In *PCE-INICÍA*, the boundary for failing was set to

50% of the maximum score. Intuitively this feels high when thinking about the standard evaluation of the Master's theses in university; if the minimum scores are met in all criteria, the work is passed. Here not only the minimum but half of the possible scores in all criteria has to be met. If using the criterion of -1.5 standard points as the benchmark, somewhat 20% of the total score should have been reached in order to be above the exceptionally low group in the *PCE-INICIA*. In *PCD-Básica*, one needs to reach 59% of the total score in order to be "Sufficient", in *PCD-Biología*, *-Historia*, and *-Parvularía* 60%, in *-Física* 63%, in *-Matemática* and *-Química* 65%, and in *-Lenguaje* as high as 68% (Table 15). Hence, the requirements for being "sufficient" are quite high.

Table 8 : The highest values for "insufficient" in the sub-tests of *INICIA*

Sub-Test	the highest score for "insufficient"	Maximum	Score/Maximum*100
PCE-INICIA	17 (not passed)	36	47.2
PCP-Básica	30	50	60.0
PCP Parvularía	30	50	60.0
PCP-Media	30	50	60.0
PCD Básica	46	80	57.5
PCD-Biología	35	60	58.3
PCD Física	37	60	61.7
PCD Matemática	38	60	63.3
PCD Química	38	60	63.3
PCD-Historia	35	60	58.3
PCD-Lenguaje	40	60	66.7
PCD-Parvularia	35	60	58.3

154. Another obvious note, derived from Appendix 4, is that the range from "insufficient" to "outstanding" varies remarkably and in some cases it is quite narrow, even too narrow. For example, in the *PCD-Biología*- and *PCP-Parvularía* tests, only six points differentiate the "insufficient" and "outstanding" test-takers which equal with 10% and 12% of the maximum score. When remembering that the reliabilities were quite low in many tests (see Table 1) and, hence, the standard errors of the measurement are high (Table 16) and the ranges seem too narrow to make the difference between the test-takers. On another day, a test taker at the upper boundary of "insufficiency" could be labelled as "outstanding" in the tests of PCP-Parvularia, PCD-Biología, and maybe also in PCD-Parvularia (see discussion about the estimation of the error in the score in Section 5.2.3).⁹¹ Practically speaking, the impression comes that the labelling system is not coherent over the tests and it is not appropriate in a high stake testing.

⁹¹ The more modern thinking of the confidential intervals (CI) would give less wide boundaries than this classical one because it takes into account the numerous in the dataset. The classical S.E.M. gives CIs of a kind but the interval is wider and it is independent of the sample size.

Table 9 : Cut-offs for and Ranges between “insufficient” and “exceptional” in the sub-tests of *INICÍA*

Sub-Test	the highest score for “insufficient”	the lowest score for “outstanding”	Range	Maximum	Range/ Maximum*100	S.E.M*
PCE- <i>INICÍA</i>	17 (not passed)	-	-	36	-	±2.63
PCP-Básica	30	39	9	50	18.0	±3.15
PCP Parvularía	30	36	6	50	12.0	±3.17
PCP-Media	30	41	11	50	22.0	±2.92
PCD Básica	46	59	13	80	16.3	±4.00
PCD-Biología	35	41	6	60	10.0	±3.54
PCD Física	37	52	15	60	25.0	±3.32
PCD Matemática	38	51	13	60	21.7	±3.35
PCD Química	38	51	13	60	21.7	±3.40
PCD-Historia	35	49	14	60	23.3	±3.25
PCD-Lenguaje	40	51	11	60	18.3	±3.92
PCD-Parvularia	35	42	7	60	11.7	±3.46

* Standard error of measurement $\sigma_E = \sigma_X \sqrt{1 - \text{Rel}}$ on the basis of Total score (see also Table 1)

155. The labelling suggested in this report would be more recommendable than what was used in *Prueba INICÍA*: to equate the test scores over the tests and to use the latent ability (Theta) as the indicator for the cut-offs rather than standardizing the scores within the single test. This causes the boundaries to be comparable over the different tests of different difficulty levels. Another question is where the cut-offs should be; because of the norm-referenced testing, no “true” or fixed cut-offs exists. The rule of “±1.5 std. units” is one option to detect the exceptionally low- and high-levelled test-takers. These boundaries and the test score values are seen in Table 17 (see also more exhaustively in Appendix 3). By using these, somewhat rougher, boundaries, it does not lead to the situation where the true abilities of the “insufficient” and “outstanding” could be the same.

Table 10 : Cut-offs of “exceptionally low”, “medium”, and “exceptionally high” suggested by the criterion of “±1.5 std. units” in the sub-tests of *INICÍA*

Sub-Test	the highest score for “exceptionally low”	medium	the lowest score for “exceptionally high”	Range	Maximum	Range/ Maximum*100
PCE- <i>INICÍA</i>	7	18	29	22	36	61,1
PCP-Básica A	15	32	44	29	50	58,0
PCP-Básica B	16	33	44	28	50	56,0
PCP Parvularía A	14	30	43	29	50	58,0
PCP Parvularía B	14	30	43	29	50	58,0
PCP-Media A	15	32	44	29	50	58,0
PCP-Media B	16	33	45	29	50	58,0
PCD Básica A	20	44	65	45	80	56,3
PCD Básica B	21	47	68	47	80	58,8
PCD-Biología	15	33	48	33	60	55,0
PCD Física	12	31	48	36	60	60,0
PCD Matemática	17	36	51	34	60	56,7
PCD Química	16	34	49	33	60	55,0
PCD-Historia A	20	40	53	33	60	55,0
PCD-Historia B	20	39	53	33	60	55,0
PCD-Lenguaje A	19	39	53	34	60	56,7
PCD-Lenguaje B	20	40	54	34	60	56,7
PCD-Parvularia A	16	36	51	35	60	58,3
PCD-Parvularia B	16	35	51	35	60	58,3

CONCLUSIONS AND SUMMARY

156. **The objective of this report was to evaluate to what extent the “Prueba INICÍA” instrument could be used for a teacher exit exam and what adaptations would be needed.** The focus in this report is in the psychometrical and validity aspects of the tests.

157. Development and Implementation: The test development methodology, piloting and the characteristics of the items and tests are reported thoroughly. The documentation is professionally done, exhaustive, and helpful for the next round of test constructors. The relevant units of universities were given the work to do. The reported procedures of the test assembly fulfill the criteria of a professionally-done work: the item writers were selected out of experienced professionals, the test assemblers were professionals, the Table of Specifications were prepared adequately, the relevant stakeholders were involved in the processes or at least they were informed of the processes, the item analysis is done by using proper and adequate practices, and the confidentiality was secured during the process.

158. Though the procedures were adequate in many ways, it seems that the selection of the sample for the piloting was most probably not very successful. The piloting sample was compiled by using volunteer students and teachers. It is known on the basis of the evaluation that there are quite many non-discriminative items. It may be possible that the reason for the low accuracy of the tests lies in the less succeeded sampling in the piloting phase. Additionally, no documentation is found of the final testing, and the related procedures. Hence, it is practically impossible to assess the data management and -analysis or scoring procedure of the final phase.

159. Validity Issues: The aim of the *INICÍA* is “to monitor the knowledge and skills of new graduates from pre-teacher training institutions”. It is quite obvious, that the tests measure the *knowledge* dimension of the new graduates and it gives only a restricted picture of the *skills* of the graduates. Such dimensions of a good teacher as the personality of the teacher, pedagogical skills in action, and classroom management are measured in lesser or nonexistent quantity.

160. From the face validity viewpoint, the tests are interesting, professional looking, and versatile though restricted to Multiple Choice type of questions. The reports describing the procedures of developing the instruments show that the work was done professionally and seriously. To make the tests even more versatile, a couple of productive items would raise the standard.

161. From the structure validity viewpoint, the structures of the tests are well-documented by the test developers, they are based on a relevant theoretical framework (school curricula), and the observed structure correspond with the aimed one. Hence, the structures of the tests seem valid. However, by maximizing the validity over the reliability may be one reason why the reliabilities of the sub-tests of *INICÍA* are quite low. The reliabilities for high stake tests are high or sufficient only in the tests of *PCD-Física* ($\alpha = 0.91$) and *PCD-Matemática* ($\alpha = 0.88$). The number of linking items is proper for the stable estimation of the items parameters over the versions.

162. From the content validity viewpoint, the contents of the tests were based on either the national curricula or the *Estándares Orientadores para Egresados de Carreras de Pedagogía en Educación Básica, Parvularia o Media*. Hence, there is no doubt that the contents of the tests are valid to measure the knowledge base of the beginning teachers. An exhaustive analysis of the contents would need quite may substance experts.

163. From the ecological validity viewpoint, the depth of the tests is versatile for testing the cognitive processes of the graduate teacher. The proportions of Knowledge-, Comprehension-, and Higher skills items were fixed to 30%, 40% and 30% respectively. The number of recall type of items feels quite high in comparison with the international practice; the international student assessment settings as PISA and TIMSS seem to be geared toward application rather than memorizing things. In *INICÍA*, the Application and Higher

skills are combined though it seems, however, that these items are geared toward Higher skills even though they are called “skill-related items”.

164. All in all, the *INICÍA* examination seems professionally made set of tests, versatile and motivating though restricted to measure the knowledge aspect of the graduating teacher. The *INICÍA* examination is very limited from some other relevant aspects of the “good teaching”, such as the classroom management, pedagogical skills, or personal traits of the graduates.

165. Pyschometric Properties: The reliabilities of the sub-tests of *INICÍA* are quite low in many cases when keeping in mind that the test is used as a high stake test. The reliability of the scores reflects strictly the accuracy and discrimination power of the test; the lower the reliability the less accurately the total score reflects the true ability of the test-takers. From this point of view, the reliabilities such as $\alpha = 0.64$ (*PCE-INICÍA*), $\alpha = 0.66$ (*PCP-Básica*), $\alpha = 0.68$ (*PCP-Parvularia*), and $\alpha = 0.69$ (*PCD-Parvularia*) are very low and $\alpha = 0.71$ (*PCD-Lenguaje*), $\alpha = 0.72$ (*PCP-Media*), $\alpha = 0.74$ (*PCD-Historia*), and $\alpha = 0.77$ (*PCD-Biología*). In many cases, the standard error of measurement is more than ± 3 points which leads to a situation in some tests that the “insufficient” and “outstanding” test-taker can be reversed.

166. Given that the tests were developed rigorously and professional, the final *INICÍA* test set includes surprising many low-discriminating items, that is, *poor* items the set of tests includes. Out of 915 items, there are 19 (2.1%) pathological items with negative item-total correlation and 294 (32.1%) of those which should have been omitted at the final phase because of very low items discrimination ($Rit < 0.20$). For the later use of the tests, it is recommendable either to omit or rewrite these to raise the standard of the tests or select new items instead of the poor and pathological ones.

167. There seems to be four kinds of challenges in the flagged items. In many cases, there is *only one alternative to select* – which happens to be the correct one. In these items, even the weakest students know, just recognize, or guess the correct answer too easily and, hence, the low item discrimination. In these cases there are also usually one or more alternatives which are never selected. It may be worthwhile to rewrite the items so that these alternatives are amended, if possible, to more attractive so that the weakest students would select those distractors. Another commonly seen challenge is that there seems to be *several “correct” answers* which attract the best students. The main law is that the best students should select the correct alternative more probable than the weaker ones. In quite many items of *INICÍA*, this does not happen. It may be worth considering revising (or at least checking) the items so that there really are not those kinds of alternatives which can be (partly) correct ones according to the latest results of the latest journals, for example. Two less common challenges are connected by the fact that the *weakest students seem to guess the correct answer too easily*. In some cases, this evidently leads to the pathological, *negative, item-test correlation*. The latter may be caused also the fact that there seems to be several items where the graphical analysis suggests that the key was not correct. Obviously, these items should be omitted or rewritten.

168. From the IRT modelling viewpoint, the difficulty levels of the items (B parameters in IRT modelling) range from $B = -4.082$ to $B = 3.14$. The distribution of the item difficulties is geared toward easier items rather than difficult items. From the test construction point of view it would be good if the really good test takers had been given an opportunity to show how good they are. Now it seems that each three most difficult item (Bio_A47, Bio_A40, and His_A40) are flagged as pathological ones; the item discrimination is negative and the percentage of correct answers is $p < 0.04$). The reason may be an incorrect key.

169. The Mantel-Haenszel statistic (MH) and a graphical evaluation were used to assess the Differential Item Functioning (DIF) of the tests. The number of cases is, in most datasets, too sparse to perform a proper DIF analysis even for the smallest number of the comparable groups, that is, when comparing two groups. However, the DIF of the items were tested on the basis of the variable *Tipo de evaluado* which has two values: 1=*Egresado de pedagogía* and 2=*Beca Vocación de Profesor o Enseña Chile*. MH gives the result as the Standard Normal distribution fractions. Statistically significant DIF would require values over 1.96.

None of 915 items showed this high value. Hence, from the statistical viewpoint, none of the items show DIF. The graphical analysis, however, shows grave discrepancies between the groups.

170. Reporting Categories: Maybe the most important question of all is whether the reporting categories (“insufficient”, “sufficient”, and “outstanding”) are fair for all test takers. It is evident that the individual tests and test versions are not at the same difficulty levels which should have been taken into account when constructing the reporting categories. Now, the mediocre test-taker with the latent ability of $\theta = 0.00$ would gain in the *PCD-Física* only 31 points while, with the same latent ability level, the test-taker in the *PCD-Historia* and in *PCD-Lenguaje* would gain 40 points even though the maximum values of the tests are the same. The latter tests are remarkably easier than the former one. Because the scores differ from each other, it would have been profitable to equate the scores before calculating the reporting categories.

171. The challenge in the reporting categories is that they are based on a set of norm-referenced tests and, hence, there are no absolute criteria where to set the boundaries for “insufficient”, “sufficient”, and “outstanding” test-taker. Then the relevant question is, who decides where the boundaries are and on what basis? In the norm-referenced testing, it may happen that *all* the candidates are good enough in an absolute sense but the norm always points out some test-takers to be the lowest ones and the others to be the highest ones. Hence, the boundaries for “insufficient”, “sufficient”, and “outstanding” are not fixed in an absolute sense.

172. The final judging of the graduate teachers, the boundaries for “insufficiency” or “failing” are relatively high. In PCE-INICÍA, the boundary for failing was set to 50% of the maximum score, in *PCD-Básica* one needs to reach 59% of the total score in order to be “Sufficient”, in *PCD-Biológica*, *-Historia*, and *-Parvularía* 60%, in *-Física* 63%, in *-Matemática* and *-Química* 65%, and in *Lenguaje* as high as 68%. Hence, the requirements for being “sufficient” are quite high. Another option, used in the studies of “weak” students, is to use the criterion of 1.5 standard points below the average as the benchmark.

173. The comparability of the standard deviations urges the equating of the test scores. This would be more recommendable than what was used in *Prueba INICÍA*. It would be better to equate the test scores over the tests and to use the latent ability (Theta) as the indicator for the cut-offs rather than standardizing the scores within the single test. Equating would cause the boundaries to be comparable over the different tests of different difficulty levels.

174. The standard errors of the measurement are high and the ranges from “insufficiency” to “outstanding” seem too narrow to make the difference between the test-takers. In another day, a test taker at the upper boundary of “insufficiency” could be labelled as “outstanding” in the tests of *PCP-Parvularia*, *PCD-Biología*, and maybe also in *PCD-Parvularia*. Practically speaking, the impression comes that the labelling system is not coherent over the tests and it is not appropriate in a high stake testing. By using the rule of “ ± 1.5 std. units” would not lead to the situation where the true abilities of the “insufficient” and “outstanding” could be the same.

175. Conclusions: From the validity viewpoint, the *INICÍA* test set is a good set of tests for the knowledge aspect of the graduating teacher: it is versatile, it looks interesting, the structures are well done and the contents seem adequate. The validity challenge comes from the ecological aspect: does the test really measure the skills needed in the real life teaching? Not necessarily; though the knowledge base of the graduate teachers is important it is not – especially at the lower grades – necessarily as important as the personal characteristics and pedagogical- and managerial skills. Adding some productive type of items would enrich the tests.

176. The technical challenge in the *INICÍA* is in low accuracy. The overall reliabilities are low for a high stake testing (in most tests, $\alpha < 0.75$). The tests include too many low-discriminating items and some pathological items. In some cases, just checking whether the key is correct may solve the problem. By omitting/rewriting the pathological and poor items would raise the standard remarkably.

177. The reporting categories are adequate but their boundaries can be criticized. The test scores should be equated and the boundaries for “insufficient”, “sufficient”, and “outstanding” should be checked. The range from insufficiency to outstanding is too narrow in some tests compared with the standard error of measurement. Another systemic of “ ± 1.5 standard units” related to equated scores could be considered; this would lead to such boundaries as “exceptionally low” and “exceptionally high”. The concept of “insufficiency” should be discussed carefully; the norm-referenced testing does not provide such indicators that could be used as benchmark for the “failing” – the labels of “failing” or “insufficient” should be used cautiously.

REFERENCES

- Angrist, J. and Pischke, J. V. (2001). "Does Teacher Training Affect Pupil Learning? Evidence From Matched Comparison in Jerusalem Public Schools." *Journal of Labor Economics*, vol. 19, #2.
- Béguin, A. (2000). *Robustness of Equating High-Stake Tests*. Enschede: Febodruk B.V.
- Berry, B., Daughtrey, A., and Wieder, A. (2010). *A better system for schools: Developing, supporting and retaining effective teachers*. New York and Hillsborough, NC: Teachers Network and the Center for Teaching Quality. Retrieved from http://teachersnetwork.org/effectiveteachers/images/CTQ_FULLLResearchReport_021810.pdf
- Betts, Julian R., Andrew Zau and Lorien Rice (2003). *Determinants of Student Achievement: New Evidence from San Diego*, San Francisco: Public Policy Institute of California.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: the classification of educational goals; Handbook I: Cognitive Domain*. New York: Longmans, Green.
- Boyd, Donald J., Pamela L. Grossman, Hamilton Lankford, Susanna Loeb, and James H. Wyckoff. 2008. "Teacher Preparation and Student Achievement." CALDER Working Paper 20. Washington, DC: The Urban Institute.
- Buddin, R. and Zamarro, G. (2009) Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics* 66(2): pp. 103-115
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2007). "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects." *Economics of Education Review* 26(6): 673–82.
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2010). "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." *Journal of Human Resources* 45(3): 655–81.
- Darling-Hammond Linda (1999a). Teacher quality and student achievement: A review of state policy evidence. Center for the Study of Teaching and Policy, University of Washington, Seattle.
- Darling-Hammond, L. (2000a). *Teacher Quality and Student Achievement*, Educational Policy Analysis Archives, Vol. 8, No. 1.
- Darling-Hammond, L. (2000b). Reforming Teacher Preparation and Licensing: Debating the Evidence," *Teachers College Record*, Vol. 102, No. 1, pp. 28-56.
- Education Internacional Latin America Regional Office (2010). Teacher Training in Latin America. Report on Case Studies in Chile, Nicaragua, Peru and the Dominican Republic. San Jose, Costa Rica.
- Ehrenberg, Ronald G. and Dominic J. Brewer (1994). *Do School and Teacher Characteristics Matter? Evidence from High School and Beyond*. *Economics of Education Review*. Vol. 13. No. I. pp. 1-17, 1994.
- Eide, E., Goldhaber, D., & Brewer, D. (2004). The Teacher Labour Market and Teacher Quality. *Oxford Review of Economic Policy*, 20, 230–244.
- Evaluación1 (2013). *Ecaluación INICÍA 2012. Informe Final. Ítem 1. Prueba de Conocimientos Disciplinarios, Educación Básica*. Universidad de Chile: Centro de Investigación Avanzada en Educación (CIAE) and Departamento de Evaluación, Medición y Registro Educational (DEMRE) & Universidad Diego Portales: Centro de Políticas Comparadas de Educación (CPCE). Santiago, 08 de Abril de 2013.
- Evaluación2 (2013). *Ecaluación INICÍA 2012. Informe Final. Ítem 1. Prueba de Conocimientos Disciplinarios, Pedagogía en Educación Media, Lenguaje y Comunicación, Historia, Geografía y Ciencias Sociales*. Universidad de Chile: Centro de Investigación Avanzada en Educación (CIAE) and Departamento de Evaluación, Medición y Registro Educational (DEMRE) & Universidad Diego Portales: Centro de Políticas Comparadas de Educación (CPCE). Santiago, 08 de Abril de 2013.
- Evaluación3 (2013). *Ecaluación INICÍA 2012. Informe Final. Ítem 3. Prueba de Conocimientos Disciplinarios, Pedagogía en Educación Media, Matemática, Biología, Física y Química*. Universidad de Chile: Centro de Investigación Avanzada en Educación (CIAE) and Departamento de Evaluación, Medición y Registro Educational (DEMRE) & Universidad Diego Portales: Centro de Políticas Comparadas de Educación (CPCE). Santiago, 08 de Abril de 2013.

- Evaluación4 (2013). *Ecaluación INICÍA 2012. Pruebas de Conocimientos Pedagógicos para Egresados de Pedagogía en Educación Parvularia, Básica y Media*. Centro Medición. Escuela de Psicología, Pontificia Universidad Católica de Chile.
- Ferguson, R. F., & Ladd, H. F. (1996). How and why money matters: An analysis of Alabama schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education*, pp. 265–298. Washington, DC: The Brookings Institution.
- Gladwell, Malcolm (2008). “Most Likely To Succeed – How do we hire when we can’t tell who’s right for the job?” *The New Yorker*, December 15, 2008.
- Goldhaber, Dan D. and Dominic J. Brewer (2000). Does teacher certification matter? High school certification status and student achievement. *Educational Evaluation and Policy Analysis* 22:129–146.
- Gordon, Robert, Thomas J. Kane and Douglas O. Staiger (2006). *The Hamilton Project. Identifying Effective Teachers Using Performance on the Job*. The Brookings Institution.
- Green, Elizabeth (2014). *Building a Better Teacher How Teaching Works (and How to Teach It to Everyone)*. W.W. Norton and Company, Inc.
- Greenberg, Julie, Arthur McKee and Kate Walsh (2013). *Teacher Prep Review. A review of the nation’s teacher preparation programs*. National Council on Teacher Quality.
- Guimarães, Raquel Rangel de Meireles and Martin Carnoy (2012). *Does Teacher Qualification Influence Student Achievement Gains? The Case of Plano de Desenvolvimento da Escola Schools in Brazil*. Mimeo.
- Gulliksen, H. 1950/1987. *Theory of Mental Tests*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Hambleton, R. K. (1993). Principles and Selected Applications of Item Response Theory. In RL Linn (Ed.), *Educational Measurement*. 3rd Ed. American Council of Education. Series of Higher Education. Oryx Press.
- Hanushek, Eric A. (1986). The Economics of Schooling. Production and Efficiency in Public Schools. *Journal of Economic Literature*. Vol. XXIV (September 1986), pp. 1141-1177.
- Hanushek, Eric A. (2011). The economic value of higher teacher quality. *Economics of Education Review* 30 (2011) 466–479.
- Hanushek, Eric A., and Dennis D. Kimko, (2000). "Schooling, Labor-Force Quality, and the Growth of Nations." *American Economic Review*, 90(5): 1184-1208.
- Hanushek, Eric A. and Rivkin S. G. (2006) Teacher Quality. In *Handbook of the Economics of Education, Volume 2, Amsterdam: North Holland*, 2006, pp. 1052–1078.
- Hanushek, Eric A. and Rivkin S. G. (2009). Harming the best: How schools affect the black-white achievement gap. *Journal of Policy Analysis and Management*. 28(3), 366–393.
- Hanushek, Eric A. and Ludger Woessmann (2008). The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature* 2008, 46:3, 607–668.
- Hanushek, Eric A. and Ludger Woessmann (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth* 2012, 17:267–321.
- Hargreaves, L. (2009) The Status and Prestige of Teachers and Teaching., In L.J. Saha, A.G. Dworkin (Eds). In *International Handbook of Research on Teachers and Teaching* (pp217–229) . Springer Science and Business Media LLC 2009
- Harris, Douglas N. and Sass, Tim R. (2009). The effects of NBPTS-certified teachers on student achievement. National Center for Analysis of Longitudinal Data in Education Research.
- Hattie, John (2003). Teachers Make a Difference: What is the research evidence? A paper presented at the Australian Council for Educational Research Annual Conference on Building Teacher Quality.
- Heuvelmans, T. (1998). *Tiaplus. Users Manual*. 1998 – 2011, M. & R. Department, Cito, Arnhem. The Netherlands.
- Jacob, B. and Lefgren, L. (2004). “Remedial Education and Student Achievement: A Regression-Discontinuity Analysis.” *Review of Economics and Statistics*. LXXXVI (1): 226-244.
- Kansanen, P. (2003). Teacher education on Finland: current models and new developments. In M. Moon, L. Vlăsceanu, & C. Barrows (Eds.), *Institutional Approaches to teacher Education Within Higher Education in Europe: Current Models and New Developments*. Bucharest: Unesco-Cepes, 85–108.
- Koljatic, Mladen and Mónica Silva (2013). *Transparencia En Evaluaciones E Instrumentos De Medición De Altas Consecuencias*”. Una publicación de Fundación Pro Acceso, Santiago, Chile.

- Lemov, Doug (2012). *Teach Like a Champion: 49 Techniques that Put Students on the Path to College (K-12)*. Wiley Inc. April 2010.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum Associates, Inc. Publishers.
- Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass: Addison-Wesley Publishing Company.
- Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29(1), Value-Added Assessment Special Issue, 67-101.
- Meckes, L., Taut, S., Bascopé, M., Valencia, E., & Manzi, J. (2012). *INICÍA* and the responses of teacher education institutions to increased accountability in Chile. Presentation at the Segundo Congreso Interdisciplinario de Investigación en Educación. Santiago, Chile, August 24th 2012. Retrieved from www.ciie2012.cl/download.php?file=sesiones/67.pdf.
- Metfessel, N., Michael, W. B., & Kirsner, D. A. (1969). Instrumentation of Bloom's and Krathwohl's taxonomies for the Writing of behavioral Objectives. *Psychology in the Schools*. 6, 227–231.
- Metzler, J. & Woessman, L. (2010). The impact of teacher knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics*, 99(2), 486–496.
- Metsämuuronen, J. (2009). *Methods assisting the assessment*. Oppimistulosten arviointi 1/2009. The Finnish National Board of Education. Helsinki: Yliopistopaino. [In Finnish].
- Metsämuuronen, J. (2013). *Handbook of Research Methods in Human Sciences*. International Methelp Oy. Available as E-book.
- Metsämuuronen TM & Metsämuuronen J (2013a). A Comparison of Nepalese and Finnish Teachers' Perceptions of Good Teaching. *Asian Journal of Humanities and Social Sciences (AJHSS)*. 1(2), August. <http://www.ajhss.org/pdfs-1/Comparison%20of%20Nepalese%20and%20Finnish.....pdf>
- Metsämuuronen TM & Metsämuuronen J (2013b). A Good Teacher – A Comparison of Nepalese and Finnish Teachers' Perceptions of What Constitutes a Good Teacher. In J Metsämuuronen & BR Kafle (eds). *Where Are We Now? Student achievement in Mathematics, Nepali and Social Studies in 2011*. Ministry of Education, Kathmandu, Nepal. 259 – 280.
- Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review*, 13(2), 125–145.
- Muijs, D. & Reynolds, D. (2005). *Effective teaching – evidence and practice*. Gateshead: Atheneum Press Ltd.
- Mullis, I. V. S. & Martin, M. O. (2011). 2011 Item Writing Guidelines. IEA, TIMSS & Pirls International Study Center. Lynch School of Education, Boston College.
- Murnane, Richard J. and Phillips, Barbara R., 1981. "Learning by doing, vintage, and selection: Three pieces of the puzzle relating teaching experience and teaching performance," *Economics of Education Review*, Elsevier, vol. 1(4), pages 453-465, August.
- Niemi, H. (2010). Teachers as high level professionals – What does it mean in teacher education? Perspectives from the Finnish teacher education. In K. G. Karras & C. C. Wolhuter (Eds.), *International Handbook of Teacher Education: Issues and Challenges* (Vol. 1 & II, pp. 237–254). Athens Greece: Atrapos.
- Niemi, H. (2011). Educating student teachers to become high quality professionals – A Finnish case. *Center for Educational Policy Studies Journal*, 1(1), 43–66.
- Niemi, H. & Jakku-Sihvonen, R. (2006). Research-based teacher education in Finland. In R. Jakku-Sihvonen & H. Niemi (Eds.), *Research-Based Teacher Education in Finland – Reflections by Finnish Teacher Educators* (pp. 31–51). Turku: Finnish Educational Research Association.
- Niemi, H. & Jakku-Sihvonen, R. (2011). Teacher education in Finland. In M. Valenčič Zuljan & J. Vogrinc (Eds.), *European Dimensions of Teacher Education: Similarities and Differences* (pp. 33–51). Slovenia: University of Ljubljana & The National School of Leadership in Education.
- Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges (2004). *How large are teacher effects?* *Educational Evaluation and Policy Analysis*, 26(3), 237–257. See the following URL. <http://www.sesp.northwestern.edu/docs/publications/169468047044fcbd1360b55.pdf>
- PISA (2006). PISA Released Items – Reading. OECD

- PISA (2009). Take the Test. Sample Questions from OECD's PISA Assessments. OECD.
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks pædagogiske Institut. Studies in Mathematical Psychology I. Copenhagen: Nielsen & Lydiche.
- Rivkin Steven G., Eric A. Hanushek, and John F. Kain (2005). Teachers, Schools, and Academic Achievement. *Econometrics*, 73(2), 417–458.
- Räsänen, P. & Närhi, V. (2013). Heikkojen oppijoiden koulupolku. [The path of the weak students]. In J. Metsämuuronen (Ed.), *Perusopetuksen matematiikan oppimistulosten pitkittäisarviointi vuosina 2005–2012*. [Longitudinal analysis of the Mathematical Achievement in the Compulsory Education in 2005–2012]. Koulutuksen seurantaraportit 2013:4. Opetushallitus. Tampere: Juvenes Print – Suomen Yliopistopaino Oy. 173–230. [In Finnish]
- Räsänen, P., Närhi, V. & Aunio, P. (2010). Matematiikassa heikosti suoriutuvat oppilaat perusopetuksen 6. luokan alussa. [The weak students in Mathematics at the beginning of the 6th grade] In E. K. Niemi & J. Metsämuuronen (Eds.), *Miten matematiikan taidot kehittyvät? Matematiikan oppimistulokset peruskoulun viidennen vuosiluokan jälkeen vuonna 2008*. [How the Mathematical skills are developing? Achievement in Mathematics in Compulsory Education after 5th grade in the year 2008]. Koulutuksen seurantaraportit 2010:2. Helsinki: Opetushallitus. [In Finnish]
- Ravitch, Diane (2013). *Reign of Error. The Hoax of the Privatization Movement and the Danger to America's Public Schools*. Random House, 2013.
- Sahlberg, P. (2011a). The Professional Educator: Lessons from Finland. *American Educator* 35(2), 34–38.
- Sahlberg, P. (2011b). Lessons from Finland: Where the Country's Education System Rose to the Top in Just a Couple Decades. *Education Digest*, 77(3), 18–24.
- Schleicher, A. (2011). Is the Sky the Limit to Education Improvement? *Phi Delta Kappan*, 93(2), 58–63.
- Stewart, Vivian (2012). *A World-class Education: Learning from International Models of Excellence and Innovation*. ASCD, 2012.
- TIMSS (2007). TIMSS 2003 Science Items. Released Set, Eight grade. IEA, TIMSS & Pirls International Study Center. Lynch School of Education, Boston College.
- TIMSS (2009a). TIMSS 2007 User Guide for the international Database. Released Items, Mathematics – Eight grade. IEA, TIMSS & Pirls International Study Center. Lynch School of Education, Boston College.
- TIMSS (2009b). TIMSS 2007 User Guide for the international Database. Released Items, Science – Eight grade. IEA, TIMSS & Pirls International Study Center. Lynch School of Education, Boston College.
- Verhelst ND, Glas CAW, Verstralen HHFM (1995). *One-Parameter Logistic Model OPLM*. Cito, Arnhem.
- Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on testing Problems*. Educational Testing Service, Princeton, NJ.

APPENDIX A

ITEM PARAMETERS OF THE ITEMS IN *INICIA*

Table A.1A⁹² Item parameters of *PCD-Básica*

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
BAS_AD01	-2,339	0,139	0,91	0,02	poor
BAS_AD02	-1,945	0,120	0,88	0,09	poor
BAS_AD03	-0,957	0,126	0,74	0,35	OK
BAS_AD04	-1,710	0,110	0,85	0,19	poor
BAS_AD05	-0,272	0,114	0,59	0,09	poor
BAS_AD06	-0,341	0,081	0,60	0,24	OK
BAS_AD07	0,596	0,082	0,38	0,27	OK
BAS_AD08	0,889	0,119	0,32	0,23	OK
BAS_AD09	1,298	0,129	0,25	0,24	OK
BAS_AD10	-1,776	0,160	0,86	0,28	OK
BAS_AD11	-1,722	0,111	0,85	0,34	OK
BAS_AD12	-1,269	0,097	0,79	0,19	OK
BAS_AD13	-0,285	0,114	0,60	0,37	OK
BAS_AD14	-0,374	0,082	0,61	0,28	OK
BAS_AD15	-1,127	0,093	0,76	0,30	OK
BAS_AD16	-0,224	0,081	0,57	0,30	OK
BAS_AD17	-1,631	0,152	0,84	0,15	poor
BAS_AD18	-0,845	0,088	0,71	0,27	OK
BAS_AD19	-1,191	0,134	0,78	0,32	OK
BAS_AD20	0,989	0,121	0,30	0,34	OK
BAS_AD21	-0,046	0,112	0,54	0,35	OK
BAS_AD22	-1,444	0,101	0,81	0,29	OK
BAS_AD23	-0,819	0,123	0,71	0,25	OK
BAS_AD24	-1,300	0,138	0,80	0,34	OK
BAS_AD25	-0,910	0,125	0,73	0,36	OK
BAS_AD26	0,389	0,113	0,44	0,10	poor
BAS_AD27	0,778	0,117	0,35	0,37	OK
BAS_AD28	0,610	0,082	0,38	0,35	OK
BAS_AD29	0,888	0,086	0,32	0,33	OK
BAS_AD30	-0,501	0,083	0,64	0,31	OK
BAS_AD31	-0,481	0,116	0,64	0,37	OK
BAS_AD32	-0,083	0,080	0,54	0,23	OK
BAS_AD33	0,452	0,113	0,42	0,24	OK
BAS_AD34	1,185	0,126	0,27	0,11	poor
BAS_AD35	-0,234	0,113	0,58	0,24	OK
BAS_AD36	0,889	0,119	0,32	0,34	OK
BAS_AD37	0,619	0,115	0,38	0,20	OK
BAS_AD38	-1,801	0,161	0,86	0,23	OK
BAS_AD39	-1,437	0,143	0,82	0,44	OK
BAS_AD40	0,859	0,085	0,32	0,27	OK

⁹² The second capitalized letters - A and B - refers to Version A and Version B respectively.

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag
BAS_AD41	-1,572	0,105	0,83	0,24	OK
BAS_AD42	0,910	0,086	0,31	0,19	OK
BAS_AD43	0,616	0,082	0,38	0,08	poor
BAS_AD44	-1,156	0,132	0,77	0,24	OK
BAS_AD45	-1,654	0,153	0,85	0,14	poor
BAS_AD46	-0,590	0,118	0,66	0,20	OK
BAS_AD47	0,132	0,080	0,49	0,22	OK
BAS_AD48	0,503	0,114	0,41	0,22	OK
BAS_AD49	-1,397	0,141	0,81	0,29	OK
BAS_AD50	1,201	0,126	0,26	0,16	poor
BAS_AD51	-0,454	0,082	0,62	0,31	OK
BAS_AD52	0,164	0,080	0,48	0,28	OK
BAS_AD53	0,778	0,117	0,35	0,25	OK
BAS_AD54	-0,973	0,127	0,74	0,24	OK
BAS_AD55	-1,631	0,152	0,84	0,12	poor
BAS_AD56	-0,941	0,126	0,74	0,09	poor
BAS_AD57	0,875	0,119	0,33	0,17	poor
BAS_AD58	-0,819	0,123	0,71	0,24	OK
BAS_AD59	-0,147	0,080	0,55	0,30	OK
BAS_AD60	0,351	0,112	0,45	0,34	OK
BAS_AD61	-1,414	0,101	0,81	0,10	poor
BAS_AD62	0,760	0,084	0,34	0,20	OK
BAS_AD63	1,130	0,090	0,27	0,25	OK
BAS_AD64	2,463	0,188	0,09	0,14	poor
BAS_AD65	-0,864	0,124	0,72	0,23	OK
BAS_AD66	0,078	0,112	0,51	0,22	OK
BAS_AD67	-0,789	0,122	0,71	0,21	OK
BAS_AD68	0,252	0,112	0,47	0,12	poor
BAS_AD69	0,316	0,080	0,44	0,20	OK
BAS_AD70	0,833	0,118	0,34	0,20	OK
BAS_AD71	0,402	0,113	0,43	0,31	OK
BAS_AD72	0,819	0,118	0,34	0,18	poor
BAS_AD73	0,910	0,086	0,31	0,23	OK
BAS_AD74	-0,033	0,112	0,54	0,27	OK
BAS_AD75	1,970	0,157	0,15	0,03	poor
BAS_AD76	-1,038	0,129	0,75	0,15	poor
BAS_AD77	-0,334	0,081	0,60	0,17	poor
BAS_AD78	-0,864	0,124	0,72	0,24	OK
BAS_AD79	-0,045	0,080	0,53	0,30	OK
BAS_AD80	-0,221	0,113	0,58	0,22	OK

Table A.1B Item parameters of *PCD-Básica* (omitted the linking items)

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag
BAS_BD03	-0,966	0,127	0,73	0,20	OK
BAS_BD05	-0,440	0,117	0,61	0,26	OK
BAS_BD08	-0,663	0,121	0,66	0,28	OK
BAS_BD09	0,123	0,114	0,48	0,10	poor
BAS_BD10	-1,768	0,159	0,85	0,18	poor
BAS_BD13	0,411	0,116	0,41	0,13	poor
BAS_BD17	-0,780	0,123	0,69	0,14	poor
BAS_BD19	-0,453	0,117	0,62	0,23	OK
BAS_BD20	0,683	0,119	0,35	0,18	poor
BAS_BD21	-0,692	0,121	0,67	0,36	OK
BAS_BD23	-0,467	0,117	0,62	0,44	OK
BAS_BD24	-1,081	0,131	0,75	0,26	OK
BAS_BD25	-0,239	0,115	0,57	0,33	OK
BAS_BD26	-0,07	0,114	0,53	0,27	OK
BAS_BD27	-1,768	0,159	0,85	0,30	OK
BAS_BD31	-0,467	0,117	0,62	0,33	OK
BAS_BD33	0,292	0,115	0,44	0,31	OK
BAS_BD34	0,827	0,122	0,32	0,13	poor
BAS_BD35	-1,352	0,140	0,79	0,40	OK
BAS_BD36	-0,319	0,116	0,59	0,23	OK
BAS_BD37	0,227	0,115	0,46	0,41	OK
BAS_BD38	0,451	0,116	0,40	0,35	OK
BAS_BD39	0,887	0,123	0,31	0,15	poor
BAS_BD44	-2,388	0,200	0,91	0,26	OK
BAS_BD45	-0,266	0,115	0,57	0,20	OK
BAS_BD46	-0,663	0,121	0,66	0,31	OK
BAS_BD48	-0,480	0,118	0,62	0,38	OK
BAS_BD49	-2,073	0,177	0,89	0,28	OK
BAS_BD50	0,726	0,120	0,34	0,38	OK
BAS_BD53	0,007	0,114	0,51	0,11	poor
BAS_BD54	-2,013	0,173	0,88	0,32	OK
BAS_BD55	-0,522	0,118	0,63	0,18	poor
BAS_BD56	-0,536	0,118	0,64	0,15	poor
BAS_BD57	0,504	0,117	0,39	0,21	OK
BAS_BD58	-1,257	0,136	0,78	0,28	OK
BAS_BD60	-0,359	0,116	0,60	0,24	OK
BAS_BD64	0,504	0,117	0,39	0,11	poor
BAS_BD65	-0,998	0,128	0,73	0,32	OK
BAS_BD66	-0,950	0,127	0,72	0,23	OK
BAS_BD67	-0,550	0,119	0,64	0,19	poor
BAS_BD68	-0,750	0,122	0,68	0,24	OK
BAS_BD70	-0,359	0,116	0,60	0,34	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag
BAS_BD71	0,240	0,115	0,45	0,25	OK
BAS_BD72	0,641	0,119	0,36	0,21	OK
BAS_BD74	-0,083	0,114	0,53	0,30	OK
BAS_BD75	1,241	0,133	0,24	0,15	poor
BAS_BD76	0,437	0,116	0,41	0,17	poor
BAS_BD78	-0,174	0,115	0,55	0,15	poor
BAS_BD80	-1,954	0,169	0,87	0,10	poor

Table A.2A Item parameters of PCP-Básica

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
BAS_AP01	0,605	0,117	0,386	0,250	OK
BAS_AP02	0,419	0,115	0,429	0,159	Poor
BAS_AP03	1,931	0,159	0,148	0,063	Poor
BAS_AP04	-0,546	0,120	0,654	0,266	OK
BAS_AP05	-0,798	0,125	0,707	0,075	Poor
BAS_AP06	-1,112	0,134	0,765	0,201	OK
BAS_AP07	0,619	0,117	0,383	0,131	Poor
BAS_AP08	0,458	0,116	0,420	0,237	OK
BAS_AP09	-1,553	0,152	0,833	0,298	OK
BAS_AP10	0,223	0,115	0,475	0,207	OK
BAS_AP11	-0,622	0,084	0,665	0,298	OK
BAS_AP12	-1,185	0,137	0,778	0,166	Poor
BAS_AP13	0,416	0,081	0,422	0,241	OK
BAS_AP14	-0,677	0,123	0,682	0,190	Poor
BAS_AP15	-0,909	0,128	0,728	0,224	OK
BAS_AP16	-0,914	0,088	0,722	0,143	Poor
BAS_AP17	-1,311	0,097	0,793	0,116	Poor
BAS_AP18	-0,263	0,081	0,581	0,134	Poor
BAS_AP19	-1,774	0,163	0,861	0,153	Poor
BAS_AP20	0,565	0,117	0,395	0,111	Poor
BAS_AP21	-2,514	0,148	0,926	0,198	OK
BAS_AP22	-0,622	0,084	0,662	0,128	Poor
BAS_AP23	-0,860	0,087	0,712	0,259	OK
BAS_AP24	-2,800	0,244	0,944	0,129	Poor
BAS_AP25	-2,100	0,184	0,895	0,388	OK
BAS_AP26	-0,692	0,123	0,685	0,297	OK
BAS_AP27	-1,090	0,092	0,755	0,276	OK
BAS_AP28	-0,860	0,087	0,711	0,229	OK
BAS_AP29	-1,368	0,099	0,802	0,263	OK
BAS_AP30	-0,263	0,081	0,581	0,228	OK
BAS_AP31	-0,113	0,115	0,556	0,398	OK
BAS_AP32	-0,490	0,119	0,642	0,267	OK
BAS_AP33	-0,845	0,126	0,716	0,302	OK
BAS_AP34	-0,942	0,129	0,735	0,145	Poor
BAS_AP35	-1,734	0,111	0,853	0,215	OK
BAS_AP36	0,068	0,114	0,512	0,207	OK
BAS_AP37	0,146	0,114	0,494	0,395	OK
BAS_AP38	-1,016	0,090	0,741	0,161	Poor
BAS_AP39	0,830	0,085	0,328	0,150	Poor
BAS_AP40	-0,100	0,115	0,552	0,290	OK
BAS_AP41	-2,205	0,191	0,904	0,102	Poor

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
BAS_AP42	-0,937	0,089	0,726	0,188	Poor
BAS_AP43	-0,942	0,129	0,735	0,249	OK
BAS_AP44	-0,036	0,080	0,529	0,264	OK
BAS_AP45	-0,721	0,085	0,682	0,189	Poor
BAS_AP46	-0,285	0,117	0,596	0,136	Poor
BAS_AP47	-1,112	0,134	0,765	0,136	Poor
BAS_AP48	-1,573	0,105	0,832	0,302	OK
BAS_AP49	0,017	0,115	0,525	0,141	Poor
BAS_AP50	0,198	0,114	0,481	0,250	OK

Table A.2B PCP-Básica (omitted the linking items)

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
BAS_BP01	0,084	0,111	0,490	0,262	OK
BAS_BP02	0,157	0,111	0,472	0,183	Poor
BAS_BP03	0,439	0,113	0,406	0,100	Poor
BAS_BP04	-0,706	0,118	0,672	0,348	OK
BAS_BP05	-0,652	0,117	0,661	0,336	OK
BAS_BP06	-1,451	0,139	0,809	0,199	OK
BAS_BP07	-0,904	0,122	0,713	0,139	Poor
BAS_BP08	0,591	0,114	0,371	0,221	OK
BAS_BP09	-1,086	0,127	0,748	0,306	OK
BAS_BP11	-0,456	0,114	0,617	0,260	OK
BAS_BP12	-0,612	0,116	0,652	0,430	OK
BAS_BP14	-1,358	0,136	0,794	0,187	Poor
BAS_BP16	-0,599	0,116	0,649	0,223	OK
BAS_BP20	-1,592	0,145	0,829	0,048	Poor
BAS_BP21	0,696	0,116	0,348	0,140	Poor
BAS_BP24	-1,134	0,128	0,757	0,318	OK
BAS_BP25	-2,343	0,190	0,910	0,178	Poor
BAS_BP27	0,427	0,113	0,409	0,290	OK
BAS_BP31	-1,376	0,136	0,797	0,097	Poor
BAS_BP32	-2,343	0,190	0,910	0,201	OK
BAS_BP33	-0,533	0,115	0,635	0,094	Poor
BAS_BP34	1,187	0,127	0,249	0,255	OK
BAS_BP35	-0,761	0,119	0,684	0,188	Poor
BAS_BP36	-0,231	0,112	0,565	0,299	OK
BAS_BP40	0,489	0,113	0,394	0,192	OK
BAS_BP41	0,736	0,117	0,339	0,347	OK
BAS_BP42	-1,656	0,148	0,838	0,142	Poor
BAS_BP45	-2,273	0,185	0,904	0,281	OK
BAS_BP48	-1,998	0,167	0,878	0,067	Poor
BAS_BP49	-0,097	0,111	0,533	0,168	Poor
BAS_BP50	0,818	0,118	0,322	0,134	Poor

Table A.3 Item parameters of *PCD-Biologia*

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
BIO_AD01	-1,345	0,285	0,800	0,235	OK
BIO_AD02	-0,667	0,245	0,675	0,215	OK
BIO_AD03	-1,118	0,268	0,762	0,061	Poor
BIO_AD04	-1,118	0,268	0,762	0,183	Poor
BIO_AD05	-0,980	0,260	0,738	0,453	OK
BIO_AD06	1,299	0,268	0,237	0,201	OK
BIO_AD07	0,248	0,230	0,463	0,469	OK
BIO_AD08	0,353	0,231	0,438	0,380	OK
BIO_AD09	0,406	0,232	0,425	0,425	OK
BIO_AD10	0,459	0,233	0,412	0,033	Poor
BIO_AD11	-1,807	0,330	0,863	0,102	Poor
BIO_AD12	-0,980	0,260	0,738	0,246	OK
BIO_AD13	-0,275	0,233	0,588	0,265	OK
BIO_AD14	-0,980	0,260	0,738	0,407	OK
BIO_AD15	-0,064	0,230	0,537	0,359	OK
BIO_AD16	-0,494	0,238	0,637	0,151	Poor
BIO_AD17	-0,850	0,253	0,713	0,373	OK
BIO_AD18	1,162	0,259	0,263	0,189	Poor
BIO_AD19	0,622	0,237	0,375	0,454	OK
BIO_AD20	1,447	0,278	0,212	0,138	Poor
BIO_AD21	1,033	0,252	0,287	0,203	OK
BIO_AD22	0,406	0,232	0,425	0,159	Poor
BIO_AD23	0,459	0,233	0,412	0,292	OK
BIO_AD24	-2,042	0,359	0,887	0,309	OK
BIO_AD25	-1,703	0,318	0,850	0,297	OK
BIO_AD26	-0,788	0,250	0,700	0,163	Poor
BIO_AD27	-0,850	0,253	0,713	0,421	OK
BIO_AD28	-1,345	0,285	0,800	0,279	OK
BIO_AD29	-0,116	0,231	0,550	0,296	OK
BIO_AD30	1,096	0,256	0,275	0,410	OK
BIO_AD31	0,792	0,242	0,338	0,137	Poor
BIO_AD32	0,850	0,244	0,325	0,274	OK
BIO_AD33	1,607	0,291	0,188	0,000	Poor
BIO_AD34	1,299	0,268	0,237	0,072	Poor
BIO_AD35	-3,660	0,719	0,975	-0,001	Pathological
BIO_AD36	-1,807	0,330	0,863	0,031	Poor
BIO_AD37	-1,427	0,292	0,812	0,238	OK
BIO_AD38	-1,191	0,273	0,775	0,281	OK
BIO_AD39	-0,275	0,233	0,588	0,138	Poor
BIO_AD40	3,410	0,591	0,037	-0,075	Pathological

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
BIO_AD41	0,092	0,229	0,500	0,384	OK
BIO_AD42	0,301	0,231	0,450	0,420	OK
BIO_AD43	-2,042	0,359	0,887	0,272	OK
BIO_AD44	-1,514	0,300	0,825	0,218	OK
BIO_AD45	-1,266	0,279	0,787	-0,007	Pathological
BIO_AD46	-0,667	0,245	0,675	0,221	OK
BIO_AD47	3,410	0,591	0,037	-0,098	Pathological
BIO_AD48	0,144	0,230	0,487	0,306	OK
BIO_AD49	-0,608	0,242	0,662	0,402	OK
BIO_AD50	-0,384	0,235	0,613	0,072	Poor
BIO_AD51	0,353	0,231	0,438	0,310	OK
BIO_AD52	0,144	0,230	0,487	-0,135	Pathological
BIO_AD53	2,097	0,342	0,125	0,182	Poor
BIO_AD54	-2,327	0,400	0,912	0,280	OK
BIO_AD55	-0,494	0,238	0,637	0,145	Poor
BIO_AD56	-0,439	0,237	0,625	0,374	OK
BIO_AD57	0,301	0,231	0,450	0,219	OK
BIO_AD58	0,144	0,230	0,487	0,198	OK
BIO_AD59	-0,012	0,230	0,525	0,266	OK
BIO_AD60	-0,012	0,230	0,525	0,405	OK

Table A.4 Item parameters of PCD-Física

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
FYS_AD01	-0,101	0,285	0,537	0,367	OK
FYS_AD02	1,406	0,339	0,222	0,546	OK
FYS_AD03	0,383	0,288	0,426	0,568	OK
FYS_AD04	-0,429	0,291	0,611	0,599	OK
FYS_AD05	-0,687	0,300	0,667	0,397	OK
FYS_AD06	-0,778	0,305	0,685	0,488	OK
FYS_AD07	0,06	0,285	0,500	0,683	OK
FYS_AD08	-0,513	0,294	0,630	0,514	OK
FYS_AD09	-0,429	0,291	0,611	0,443	OK
FYS_AD10	-0,021	0,285	0,519	0,418	OK
FYS_AD11	-0,182	0,286	0,556	0,610	OK
FYS_AD12	-0,967	0,315	0,722	0,414	OK
FYS_AD13	-0,429	0,291	0,611	0,450	OK
FYS_AD14	0,808	0,301	0,333	0,348	OK
FYS_AD15	0,22	0,286	0,463	0,627	OK
FYS_AD16	0,14	0,285	0,481	0,473	OK
FYS_AD17	-0,101	0,285	0,537	0,422	OK
FYS_AD18	0,808	0,301	0,333	0,274	OK
FYS_AD19	-0,687	0,300	0,667	0,213	OK
FYS_AD20	-0,967	0,315	0,722	0,465	OK
FYS_AD21	-1,172	0,329	0,759	0,444	OK
FYS_AD22	-2,143	0,441	0,889	0,297	OK
FYS_AD23	1,19	0,322	0,259	0,500	OK
FYS_AD24	1,648	0,361	0,185	0,401	OK
FYS_AD25	-0,345	0,289	0,593	0,338	OK
FYS_AD26	0,14	0,285	0,481	0,379	OK
FYS_AD27	0,383	0,288	0,426	0,322	OK
FYS_AD28	-0,429	0,291	0,611	0,403	OK
FYS_AD29	-0,687	0,300	0,667	0,539	OK
FYS_AD30	0,808	0,301	0,333	0,365	OK
FYS_AD31	0,72	0,297	0,352	0,423	OK
FYS_AD32	0,899	0,305	0,315	0,531	OK
FYS_AD33	1,09	0,316	0,278	0,432	OK
FYS_AD34	0,465	0,290	0,407	0,556	OK
FYS_AD35	-0,182	0,286	0,556	0,316	OK
FYS_AD36	1,09	0,316	0,278	0,354	OK
FYS_AD37	0,14	0,285	0,481	0,406	OK
FYS_AD38	-0,513	0,294	0,630	0,129	Poor
FYS_AD39	0,14	0,285	0,481	0,160	Poor
FYS_AD40	1,782	0,375	0,167	0,345	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
FYS_AD41	-0,182	0,286	0,556	0,460	OK
FYS_AD42	-0,513	0,294	0,630	0,202	OK
FYS_AD43	-2,914	0,600	0,944	0,087	Poor
FYS_AD44	-0,345	0,289	0,593	0,097	Poor
FYS_AD45	0,06	0,285	0,500	0,321	OK
FYS_AD46	0,22	0,286	0,463	0,182	Poor
FYS_AD47	1,406	0,339	0,222	0,407	OK
FYS_AD48	0,548	0,292	0,389	0,533	OK
FYS_AD49	-0,967	0,315	0,722	0,274	OK
FYS_AD50	-0,513	0,294	0,630	0,403	OK
FYS_AD51	-0,101	0,285	0,537	0,294	OK
FYS_AD52	0,633	0,294	0,370	0,510	OK
FYS_AD53	-0,101	0,285	0,537	0,276	OK
FYS_AD54	-0,513	0,294	0,630	0,224	OK
FYS_AD55	0,548	0,292	0,389	0,258	OK
FYS_AD56	-1,962	0,414	0,870	0,224	OK
FYS_AD57	-0,429	0,291	0,611	0,409	OK
FYS_AD58	-0,021	0,285	0,519	0,479	OK
FYS_AD59	1,19	0,322	0,259	0,202	OK
FYS_AD60	0,301	0,287	0,444	0,587	OK

Table A.5 Item parameters of *PCD-Matemática*

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
MAT_AD01	-3,033	0,364	0,955	0,056	Poor
MAT_AD02	-2,692	0,314	0,939	0,274	OK
MAT_AD03	-2,598	0,302	0,933	0,062	Poor
MAT_AD04	-1,234	0,188	0,788	0,297	OK
MAT_AD05	-1,624	0,210	0,844	0,181	Poor
MAT_AD06	-0,339	0,160	0,615	0,407	OK
MAT_AD07	0,295	0,156	0,469	0,487	OK
MAT_AD08	-0,140	0,157	0,570	0,313	OK
MAT_AD09	-0,389	0,161	0,626	0,315	OK
MAT_AD10	-0,140	0,157	0,570	0,242	OK
MAT_AD11	0,953	0,166	0,324	0,274	OK
MAT_AD12	0,150	0,156	0,503	0,464	OK
MAT_AD13	-1,269	0,190	0,793	0,325	OK
MAT_AD14	-0,941	0,175	0,737	0,377	OK
MAT_AD15	-1,418	0,198	0,816	0,298	OK
MAT_AD16	-0,764	0,170	0,704	0,360	OK
MAT_AD17	-1,131	0,183	0,771	0,239	OK
MAT_AD18	-0,415	0,161	0,631	0,376	OK
MAT_AD19	-0,793	0,170	0,709	0,419	OK
MAT_AD20	0,054	0,156	0,525	0,389	OK
MAT_AD21	0,246	0,156	0,480	0,510	OK
MAT_AD22	0,690	0,161	0,380	0,381	OK
MAT_AD23	-1,581	0,208	0,838	0,181	Poor
MAT_AD24	0,271	0,156	0,475	0,513	OK
MAT_AD25	0,246	0,156	0,480	0,368	OK
MAT_AD26	-1,418	0,198	0,816	0,328	OK
MAT_AD27	-1,581	0,208	0,838	0,297	OK
MAT_AD28	-0,971	0,177	0,743	0,338	OK
MAT_AD29	-1,342	0,194	0,804	0,229	OK
MAT_AD30	-1,762	0,220	0,860	0,376	OK
MAT_AD31	0,054	0,156	0,525	0,427	OK
MAT_AD32	-2,428	0,282	0,922	0,083	Poor
MAT_AD33	-0,466	0,162	0,642	0,306	OK
MAT_AD34	0,102	0,156	0,514	0,441	OK
MAT_AD35	0,690	0,161	0,380	0,254	OK
MAT_AD36	-1,131	0,183	0,771	0,439	OK
MAT_AD37	-0,116	0,157	0,564	0,121	Poor
MAT_AD38	0,392	0,157	0,447	0,440	OK
MAT_AD39	-0,189	0,158	0,581	0,394	OK
MAT_AD40	-1,034	0,179	0,754	0,287	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
MAT_AD41	-0,091	0,157	0,559	0,366	OK
MAT_AD42	1,498	0,186	0,223	0,432	OK
MAT_AD43	-1,131	0,183	0,771	0,310	OK
MAT_AD44	-0,466	0,162	0,642	0,333	OK
MAT_AD45	-1,131	0,183	0,771	0,241	OK
MAT_AD46	-0,339	0,160	0,615	0,422	OK
MAT_AD47	1,037	0,169	0,307	0,108	Poor
MAT_AD48	1,151	0,172	0,285	0,236	OK
MAT_AD49	1,753	0,199	0,184	0,398	OK
MAT_AD50	-1,099	0,182	0,765	0,129	Poor
MAT_AD51	0,054	0,156	0,525	0,346	OK
MAT_AD52	-0,140	0,157	0,570	0,458	OK
MAT_AD53	-0,466	0,162	0,642	0,354	OK
MAT_AD54	-0,822	0,171	0,715	0,271	OK
MAT_AD55	-0,708	0,168	0,693	0,355	OK
MAT_AD56	0,030	0,156	0,531	0,455	OK
MAT_AD57	1,715	0,197	0,190	0,033	Poor
MAT_AD58	0,198	0,156	0,492	0,557	OK
MAT_AD59	0,392	0,157	0,447	0,426	OK
MAT_AD60	-0,441	0,162	0,637	0,298	OK

Table A.6 Item parameters of PCD-Quimica

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
QUI_AD01	-2,251	0,532	0,907	-0,088	Pathological
QUI_AD02	-1,593	0,422	0,837	0,076	Poor
QUI_AD03	-1,426	0,401	0,814	0,114	Poor
QUI_AD04	-0,428	0,326	0,628	0,581	OK
QUI_AD05	-0,324	0,323	0,605	0,169	Poor
QUI_AD06	-0,324	0,323	0,605	0,131	Poor
QUI_AD07	0,076	0,316	0,512	0,223	OK
QUI_AD08	1,682	0,402	0,186	0,373	OK
QUI_AD09	0,274	0,317	0,465	0,194	OK
QUI_AD10	0,274	0,317	0,465	0,262	OK
QUI_AD11	1,387	0,372	0,233	0,347	OK
QUI_AD12	2,512	0,534	0,093	0,041	Poor
QUI_AD13	0,474	0,320	0,419	0,517	OK
QUI_AD14	-0,122	0,318	0,558	0,435	OK
QUI_AD15	0,274	0,317	0,465	0,427	OK
QUI_AD16	0,68	0,327	0,372	-0,024	Pathological
QUI_AD17	-0,222	0,320	0,581	0,066	Poor
QUI_AD18	-0,122	0,318	0,558	0,582	OK
QUI_AD19	0,576	0,323	0,395	0,377	OK
QUI_AD20	1,255	0,360	0,256	0,439	OK
QUI_AD21	-1,426	0,401	0,814	0,196	OK
QUI_AD22	-1,593	0,422	0,837	0,150	Poor
QUI_AD23	-1,781	0,448	0,860	0,277	OK
QUI_AD24	0,175	0,316	0,488	0,023	Poor
QUI_AD25	-0,644	0,336	0,674	0,359	OK
QUI_AD26	-1,781	0,448	0,860	0,165	Poor
QUI_AD27	-0,876	0,350	0,721	0,189	Poor
QUI_AD28	-0,324	0,323	0,605	0,398	OK
QUI_AD29	-0,122	0,318	0,558	-0,007	Pathological
QUI_AD30	-0,122	0,318	0,558	0,338	OK
QUI_AD31	-0,023	0,317	0,535	0,406	OK
QUI_AD32	0,076	0,316	0,512	0,429	OK
QUI_AD33	0,68	0,327	0,372	0,425	OK
QUI_AD34	0,787	0,331	0,349	0,312	OK
QUI_AD35	-3,729	1,016	0,977	0,003	Poor
QUI_AD36	-1,426	0,401	0,814	0,126	Poor
QUI_AD37	-1,781	0,448	0,860	0,409	OK
QUI_AD38	-0,758	0,342	0,698	-0,018	Pathological
QUI_AD39	-1,996	0,483	0,884	0,227	OK
QUI_AD40	-0,222	0,320	0,581	0,279	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
QUI_AD41	0,474	0,320	0,419	0,207	OK
QUI_AD42	0,68	0,327	0,372	0,180	Poor
QUI_AD43	0,787	0,331	0,349	0,140	Poor
QUI_AD44	-0,122	0,318	0,558	0,173	Poor
QUI_AD45	2,255	0,485	0,116	0,122	Poor
QUI_AD46	-1,996	0,483	0,884	0,135	Poor
QUI_AD47	-1,273	0,384	0,791	0,167	Poor
QUI_AD48	-1,001	0,359	0,744	0,257	OK
QUI_AD49	0,175	0,316	0,488	0,443	OK
QUI_AD50	0,076	0,316	0,512	0,415	OK
QUI_AD51	0,274	0,317	0,465	0,235	OK
QUI_AD52	1,529	0,385	0,209	0,277	OK
QUI_AD53	-2,57	0,605	0,930	0,220	OK
QUI_AD54	-1,426	0,401	0,814	0,232	OK
QUI_AD55	-1,132	0,371	0,767	0,491	OK
QUI_AD56	-1,132	0,371	0,767	0,286	OK
QUI_AD57	0,474	0,320	0,419	0,290	OK
QUI_AD58	0,076	0,316	0,512	0,342	OK
QUI_AD59	-0,023	0,317	0,535	0,122	Poor
QUI_AD60	0,787	0,331	0,349	0,332	OK

Table A.7 A Item parameters of *PCD-Historia*

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
HIS_AD01	-2,409	0,220	0,918	0,248	OK
HIS_AD02	-0,246	0,126	0,572	0,215	OK
HIS_AD03	-0,981	0,200	0,718	0,176	Poor
HIS_AD04	-1,657	0,167	0,839	0,335	OK
HIS_AD05	-1,677	0,239	0,832	0,297	OK
HIS_AD06	-0,484	0,128	0,628	0,166	Poor
HIS_AD07	0,115	0,181	0,473	0,492	OK
HIS_AD08	0,083	0,181	0,481	0,239	OK
HIS_AD09	-1,424	0,156	0,805	0,261	OK
HIS_AD10	-0,274	0,182	0,565	0,178	Poor
HIS_AD11	-0,668	0,132	0,667	0,277	OK
HIS_AD12	-0,542	0,187	0,626	0,200	OK
HIS_AD13	-1,103	0,205	0,740	0,404	OK
HIS_AD14	-1,677	0,239	0,832	0,241	OK
HIS_AD15	-0,014	0,181	0,504	0,394	OK
HIS_AD16	0,43	0,126	0,413	0,113	Poor
HIS_AD17	-0,144	0,181	0,534	0,132	Poor
HIS_AD18	-1,416	0,221	0,794	0,168	Poor
HIS_AD19	0,509	0,127	0,395	-0,036	Pathological
HIS_AD20	-1,857	0,253	0,855	0,472	OK
HIS_AD21	-0,718	0,191	0,664	0,260	OK
HIS_AD22	-0,209	0,182	0,550	0,303	OK
HIS_AD23	-0,209	0,182	0,550	0,036	Poor
HIS_AD24	-0,791	0,193	0,679	0,286	OK
HIS_AD25	-1,516	0,228	0,809	0,344	OK
HIS_AD26	0,41	0,184	0,405	0,288	OK
HIS_AD27	-1,231	0,211	0,763	0,167	Poor
HIS_AD28	-1,922	0,259	0,863	0,275	OK
HIS_AD29	-1,276	0,213	0,771	0,280	OK
HIS_AD30	-1,416	0,221	0,794	0,277	OK
HIS_AD31	-0,137	0,125	0,547	0,235	OK
HIS_AD32	-0,209	0,182	0,550	0,190	Poor
HIS_AD33	-0,076	0,125	0,535	0,158	Poor
HIS_AD34	-1,99	0,265	0,870	0,430	OK
HIS_AD35	-1,795	0,248	0,847	0,188	Poor
HIS_AD36	-1,677	0,239	0,832	0,088	Poor
HIS_AD37	-0,274	0,182	0,565	0,187	Poor
HIS_AD38	-0,942	0,198	0,710	0,332	OK
HIS_AD39	-0,994	0,140	0,733	0,217	OK
HIS_AD40	-1,622	0,235	0,824	0,284	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
HIS_AD41	-1,772	0,174	0,855	0,176	Poor
HIS_AD42	-1,677	0,239	0,832	0,264	OK
HIS_AD43	0,214	0,124	0,464	0,170	Poor
HIS_AD44	-1,034	0,141	0,743	0,349	OK
HIS_AD45	-0,577	0,187	0,634	0,277	OK
HIS_AD46	1,606	0,158	0,186	-0,150	Pathological
HIS_AD47	-0,828	0,194	0,687	0,516	OK
HIS_AD48	-0,682	0,190	0,656	0,141	Poor
HIS_AD49	0,212	0,182	0,450	0,218	OK
HIS_AD50	-0,682	0,190	0,656	0,062	Poor
HIS_AD51	0,277	0,182	0,435	0,061	Poor
HIS_AD52	-0,865	0,196	0,695	0,326	OK
HIS_AD53	-1,021	0,201	0,725	0,336	OK
HIS_AD54	0,905	0,197	0,298	0,282	OK
HIS_AD55	-0,44	0,185	0,603	0,363	OK
HIS_AD56	-1,575	0,163	0,828	0,270	OK
HIS_AD57	-0,014	0,181	0,504	0,174	Poor
HIS_AD58	-2,981	0,392	0,947	0,149	Poor
HIS_AD59	-0,403	0,127	0,613	0,252	OK
HIS_AD60	3,329	0,458	0,038	-0,078	Pathological

Table A.7 B Item parameters of PCD-Historia (Omitted the linking items)

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
HIS_BD01	-0,462	0,178	0,639	0,362	OK
HIS_BD02	0,782	0,179	0,347	0,383	OK
HIS_BD07	-1,692	0,239	0,854	0,159	Poor
HIS_BD08	-1,384	0,217	0,812	0,285	OK
HIS_BD10	-0,19	0,173	0,576	0,046	Poor
HIS_BD12	-1,338	0,214	0,806	0,165	Poor
HIS_BD13	0,042	0,171	0,521	0,353	OK
HIS_BD14	0,186	0,171	0,486	0,339	OK
HIS_BD15	-0,37	0,176	0,618	0,301	OK
HIS_BD17	-1,007	0,196	0,750	0,374	OK
HIS_BD18	0,1	0,171	0,507	0,239	OK
HIS_BD20	-2,532	0,330	0,931	0,222	OK
HIS_BD21	-2,646	0,347	0,938	0,154	Poor
HIS_BD22	-0,34	0,175	0,611	0,317	OK
HIS_BD23	-0,4	0,176	0,625	0,214	OK
HIS_BD24	-1,811	0,249	0,868	0,334	OK
HIS_BD25	-0,687	0,184	0,688	0,029	Poor
HIS_BD26	0,129	0,171	0,500	0,373	OK
HIS_BD27	-1,085	0,200	0,764	0,149	Poor
HIS_BD28	-1,294	0,211	0,799	0,158	Poor
HIS_BD29	-0,557	0,180	0,660	0,193	OK
HIS_BD31	-2,083	0,276	0,896	0,272	OK
HIS_BD32	-0,431	0,177	0,632	0,193	OK
HIS_BD34	0,331	0,172	0,451	0,039	Poor
HIS_BD35	-0,279	0,174	0,597	0,365	OK
HIS_BD36	0,75	0,178	0,354	0,227	OK
HIS_BD37	-0,969	0,194	0,743	0,024	Poor
HIS_BD39	-0,896	0,191	0,729	0,151	Poor
HIS_BD40	-1,811	0,249	0,868	0,257	OK
HIS_BD45	-1,085	0,200	0,764	0,190	Poor
HIS_BD46	0,75	0,178	0,354	0,376	OK
HIS_BD47	-2,333	0,304	0,917	0,169	Poor
HIS_BD48	-1,874	0,255	0,875	0,348	OK
HIS_BD49	-1,046	0,198	0,757	0,280	OK
HIS_BD50	0,331	0,172	0,451	0,299	OK
HIS_BD51	0,1	0,171	0,507	0,150	Poor
HIS_BD52	-0,309	0,175	0,604	0,223	OK
HIS_BD53	-0,19	0,173	0,576	0,326	OK
HIS_BD54	-1,811	0,249	0,868	0,145	Poor
HIS_BD56	-1,046	0,198	0,757	0,255	OK
HIS_BD58	-0,19	0,173	0,576	0,254	OK
HIS_BD59	-2,333	0,304	0,917	0,148	Poor

HIS_BD60	-0,279	0,174	0,597	0,177	Poor
----------	--------	-------	-------	-------	------

Table A.8 A Item parameters of *PCD-Lenguaje*

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
LEN_AD01	-1,659	0,260	0,884	0,204	OK
LEN_AD02	-0,521	0,249	0,700	0,174	Poor
LEN_AD03	-0,771	0,197	0,771	0,414	OK
LEN_AD04	0,072	0,171	0,583	0,258	OK
LEN_AD05	-0,46	0,246	0,688	0,110	Poor
LEN_AD06	-0,46	0,246	0,688	0,317	OK
LEN_AD07	-0,174	0,236	0,625	0,154	Poor
LEN_AD08	-0,104	0,174	0,629	0,173	Poor
LEN_AD09	0,935	0,237	0,362	0,041	Poor
LEN_AD10	-0,065	0,233	0,600	0,320	OK
LEN_AD11	0,101	0,171	0,583	0,240	OK
LEN_AD12	-0,352	0,180	0,680	0,250	OK
LEN_AD13	-1,428	0,317	0,850	0,312	OK
LEN_AD14	0,405	0,229	0,487	-0,030	Pathological
LEN_AD15	0,041	0,231	0,575	0,363	OK
LEN_AD16	0,198	0,229	0,537	0,112	Poor
LEN_AD17	-2,487	0,365	0,945	0,258	OK
LEN_AD18	-1,53	0,249	0,873	0,357	OK
LEN_AD19	-1,412	0,238	0,855	0,152	Poor
LEN_AD20	-0,286	0,239	0,650	0,366	OK
LEN_AD21	0,198	0,229	0,537	0,259	OK
LEN_AD22	-0,174	0,236	0,625	0,204	OK
LEN_AD23	0,935	0,237	0,362	0,127	Poor
LEN_AD24	-2,414	0,465	0,938	-0,087	Pathological
LEN_AD25	-0,12	0,235	0,613	0,357	OK
LEN_AD26	0,094	0,231	0,562	0,366	OK
LEN_AD27	0,718	0,232	0,412	0,154	Poor
LEN_AD28	0,665	0,231	0,425	0,098	Poor
LEN_AD29	-0,848	0,267	0,762	-0,052	Pathological
LEN_AD30	0,826	0,234	0,388	0,203	OK
LEN_AD31	-0,848	0,267	0,762	0,379	OK
LEN_AD32	-0,81	0,199	0,782	0,193	OK
LEN_AD33	-2,361	0,346	0,938	0,213	OK
LEN_AD34	-1,357	0,234	0,854	0,228	OK
LEN_AD35	-1,412	0,238	0,863	0,117	Poor
LEN_AD36	-1,155	0,291	0,812	0,128	Poor
LEN_AD37	-0,45	0,184	0,699	0,294	OK
LEN_AD38	-2,487	0,365	0,947	0,095	Poor
LEN_AD39	-1,643	0,342	0,875	0,143	Poor
LEN_AD40	-0,659	0,192	0,744	0,203	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag Rit for
LEN_AD41	-0,81	0,199	0,770	0,218	OK
LEN_AD42	-0,45	0,184	0,700	0,212	OK
LEN_AD43	-0,849	0,201	0,778	0,115	Poor
LEN_AD44	-0,257	0,178	0,667	0,306	OK
LEN_AD45	-1,241	0,298	0,825	0,207	OK
LEN_AD46	-2,652	0,515	0,950	0,173	Poor
LEN_AD47	-1,017	0,210	0,805	0,281	OK
LEN_AD48	-0,712	0,259	0,738	0,252	OK
LEN_AD49	-0,771	0,197	0,767	0,123	Poor
LEN_AD50	-0,771	0,197	0,767	0,206	OK
LEN_AD51	-0,012	0,232	0,588	0,251	OK
LEN_AD52	-0,288	0,179	0,680	0,148	Poor
LEN_AD53	-4,082	1,008	0,988	0,017	Poor
LEN_AD54	-2,652	0,515	0,950	0,100	Poor
LEN_AD55	-1,332	0,307	0,838	0,346	OK
LEN_AD56	-0,995	0,278	0,787	0,303	OK
LEN_AD57	0,041	0,231	0,575	0,063	Poor
LEN_AD58	-0,521	0,249	0,700	0,369	OK
LEN_AD59	0,146	0,230	0,550	0,147	Poor
LEN_AD60	0,826	0,234	0,388	0,232	OK

Table A.8 B Item parameters of PCD-Lenguaje (Omitted the linking items)

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag Rit for
LEN_BD01	-3,73	1,009	0,985	0,055	Poor
LEN_BD02	-2,293	0,518	0,941	0,057	Poor
LEN_BD04	-0,438	0,276	0,721	0,377	OK
LEN_BD08	-0,153	0,262	0,662	0,347	OK
LEN_BD09	1,245	0,263	0,338	0,100	Poor
LEN_BD10	-1,852	0,431	0,912	0,139	Poor
LEN_BD11	-0,222	0,265	0,676	0,054	Poor
LEN_BD12	-0,438	0,276	0,721	0,069	Poor
LEN_BD13	-0,02	0,258	0,632	0,242	OK
LEN_BD14	-0,153	0,262	0,662	0,417	OK
LEN_BD16	-1,152	0,334	0,838	0,032	Poor
LEN_BD18	-1,391	0,362	0,868	0,354	OK
LEN_BD19	-0,292	0,268	0,691	0,088	Poor
LEN_BD21	-0,153	0,262	0,662	0,339	OK
LEN_BD24	0,483	0,249	0,515	0,143	Poor
LEN_BD25	0,046	0,256	0,618	0,268	OK
LEN_BD26	-0,945	0,313	0,809	0,119	Poor
LEN_BD27	0,046	0,256	0,618	0,140	Poor
LEN_BD28	-1,391	0,362	0,868	0,278	OK
LEN_BD29	-0,593	0,285	0,750	0,364	OK
LEN_BD30	-0,593	0,285	0,750	0,101	Poor
LEN_BD32	-1,528	0,380	0,882	0,340	OK
LEN_BD35	-2,293	0,518	0,941	0,229	OK
LEN_BD41	-0,851	0,305	0,794	0,137	Poor
LEN_BD45	-1,391	0,362	0,868	0,202	OK
LEN_BD50	0,98	0,255	0,397	0,149	Poor
LEN_BD51	-1,045	0,323	0,824	0,091	Poor
LEN_BD52	0,854	0,252	0,426	0,288	OK
LEN_BD53	-1,267	0,347	0,853	-0,027	Pathological
LEN_BD54	-1,68	0,403	0,897	0,164	Poor
LEN_BD55	-0,593	0,285	0,750	0,224	OK
LEN_BD56	-0,851	0,305	0,794	0,150	Poor
LEN_BD57	-0,292	0,268	0,691	0,108	Poor
LEN_BD58	0,544	0,249	0,500	0,296	OK
LEN_BD59	0,173	0,253	0,588	0,236	OK
LEN_BD60	0,236	0,252	0,574	0,270	OK

Table A.9A Item parameters of PCD-Parvularia

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
PAR_AD01	-1,851	0,204	0,851	0,146	Poor
PAR_AD02	-1,643	0,135	0,823	0,151	Poor
PAR_AD03	-2,465	0,256	0,912	0,253	OK
PAR_AD04	-1,266	0,172	0,763	0,228	OK
PAR_AD05	-1,075	0,116	0,728	0,291	OK
PAR_AD06	-0,276	0,105	0,553	0,219	OK
PAR_AD07	-1,017	0,163	0,716	0,174	Poor
PAR_AD08	-0,227	0,148	0,541	0,022	Poor
PAR_AD09	-1,102	0,117	0,733	0,296	OK
PAR_AD10	-0,265	0,105	0,550	0,251	OK
PAR_AD11	-1,387	0,178	0,784	0,088	Poor
PAR_AD12	-0,055	0,147	0,500	0,290	OK
PAR_AD13	-0,007	0,104	0,489	0,231	OK
PAR_AD14	-0,991	0,162	0,711	0,336	OK
PAR_AD15	0,306	0,106	0,414	0,144	Poor
PAR_AD16	-1,208	0,170	0,753	0,159	Poor
PAR_AD17	-1,256	0,121	0,761	0,189	Poor
PAR_AD18	-0,517	0,107	0,609	0,387	OK
PAR_AD19	-1,208	0,170	0,753	0,272	OK
PAR_AD20	-0,249	0,148	0,546	0,070	Poor
PAR_AD21	-0,055	0,147	0,500	0,135	Poor
PAR_AD22	0,269	0,149	0,423	-0,065	Pathological
PAR_AD23	0,328	0,106	0,409	0,291	OK
PAR_AD24	-0,357	0,149	0,572	0,289	OK
PAR_AD25	-1,124	0,167	0,737	0,300	OK
PAR_AD26	-1,18	0,169	0,747	0,404	OK
PAR_AD27	-0,557	0,152	0,619	0,173	Poor
PAR_AD28	-1,266	0,172	0,763	0,239	OK
PAR_AD29	-0,136	0,104	0,519	0,154	Poor
PAR_AD30	0,175	0,105	0,445	0,261	OK
PAR_AD31	-0,292	0,148	0,557	0,100	Poor
PAR_AD32	1,155	0,172	0,237	0,117	Poor
PAR_AD33	1,144	0,121	0,240	0,093	Poor
PAR_AD34	0,906	0,163	0,284	0,369	OK
PAR_AD35	1,868	0,213	0,134	0,030	Poor
PAR_AD36	-1,608	0,133	0,818	0,108	Poor
PAR_AD37	-1,18	0,169	0,747	0,345	OK
PAR_AD38	-0,94	0,160	0,701	0,351	OK
PAR_AD39	-0,914	0,160	0,696	0,320	OK
PAR_AD40	-1,048	0,116	0,723	0,199	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
PAR_AD41	0,251	0,105	0,427	0,384	OK
PAR_AD42	-0,061	0,104	0,502	0,194	OK
PAR_AD43	0,854	0,161	0,294	0,262	OK
PAR_AD44	0,932	0,164	0,278	0,278	OK
PAR_AD45	-1,483	0,182	0,799	0,161	Poor
PAR_AD46	-1,731	0,196	0,835	0,242	OK
PAR_AD47	0,313	0,150	0,412	0,052	Poor
PAR_AD48	0,074	0,148	0,469	0,304	OK
PAR_AD49	-0,314	0,148	0,562	0,114	Poor
PAR_AD50	0,601	0,109	0,348	0,329	OK
PAR_AD51	0,754	0,158	0,314	0,095	Poor
PAR_AD52	0,733	0,111	0,319	0,126	Poor
PAR_AD53	-1,451	0,181	0,794	0,029	Poor
PAR_AD54	-0,314	0,148	0,562	0,190	Poor
PAR_AD55	-0,689	0,109	0,648	0,191	OK
PAR_AD56	-0,276	0,105	0,553	0,061	Poor
PAR_AD57	-0,249	0,148	0,546	0,165	Poor
PAR_AD58	-0,791	0,156	0,670	0,200	OK
PAR_AD59	0,778	0,159	0,309	0,418	OK
PAR_AD60	1,582	0,194	0,170	0,138	Poor

Table A.9B Item parameters of *PCD-Parvularia* (Omitted the linking items)

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag Rit for
PAR_BD01	-1,335	0,175	0,774	0,039	Poor
PAR_BD03	-1,493	0,182	0,800	0,276	OK
PAR_BD04	-1,56	0,186	0,810	0,124	Poor
PAR_BD07	0,366	0,150	0,400	0,237	OK
PAR_BD08	-0,899	0,159	0,692	0,228	OK
PAR_BD11	-0,636	0,153	0,636	0,176	Poor
PAR_BD12	0,212	0,148	0,436	0,286	OK
PAR_BD14	0,126	0,148	0,456	0,089	Poor
PAR_BD16	-0,259	0,148	0,549	0,104	Poor
PAR_BD18	-0,777	0,156	0,667	0,288	OK
PAR_BD20	-0,613	0,152	0,631	0,192	OK
PAR_BD21	0,169	0,148	0,446	0,158	Poor
PAR_BD22	0,524	0,152	0,364	0,239	OK
PAR_BD24	-1,276	0,172	0,764	0,006	Poor
PAR_BD25	-1,526	0,184	0,805	0,190	Poor
PAR_BD26	-0,899	0,159	0,692	0,404	OK
PAR_BD27	-0,825	0,157	0,677	0,105	Poor
PAR_BD28	-0,924	0,160	0,697	0,213	OK
PAR_BD31	-0,777	0,156	0,667	0,289	OK
PAR_BD32	-0,195	0,147	0,533	0,321	OK
PAR_BD34	1,552	0,192	0,174	0,237	OK
PAR_BD35	0,862	0,161	0,292	0,175	Poor
PAR_BD37	-0,899	0,159	0,692	0,318	OK
PAR_BD38	-0,238	0,148	0,544	0,330	OK
PAR_BD39	-0,324	0,148	0,564	0,091	Poor
PAR_BD43	-0,825	0,157	0,677	0,310	OK
PAR_BD44	0,256	0,149	0,426	0,244	OK
PAR_BD45	-1,99	0,214	0,867	0,124	Poor
PAR_BD46	-1,397	0,178	0,785	0,296	OK
PAR_BD47	-0,238	0,148	0,544	0,365	OK
PAR_BD48	-0,39	0,149	0,579	0,281	OK
PAR_BD49	0,547	0,153	0,359	0,259	OK
PAR_BD51	0,664	0,155	0,333	0,192	OK
PAR_BD53	-2,133	0,225	0,882	0,328	OK
PAR_BD54	-0,613	0,152	0,631	0,157	Poor
PAR_BD57	0,234	0,148	0,431	0,119	Poor
PAR_BD58	0,04	0,147	0,477	0,278	OK
PAR_BD59	0,737	0,157	0,318	0,178	Poor
PAR_BD60	0,967	0,164	0,272	0,055	Poor

Table A.10A Item parameters of PCP-Parvularia

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
PAR_AP01	-1,852	0,201	0,851	0,176	Poor
PAR_AP02	-0,273	0,145	0,554	0,177	Poor
PAR_AP03	2,223	0,170	0,099	0,263	OK
PAR_AP04	0,649	0,109	0,336	0,103	Poor
PAR_AP05	0,077	0,145	0,470	0,249	OK
PAR_AP06	-0,82	0,154	0,678	0,311	OK
PAR_AP07	-0,615	0,150	0,634	0,246	OK
PAR_AP08	-1,597	0,132	0,815	0,234	OK
PAR_AP09	-0,615	0,150	0,634	0,325	OK
PAR_AP10	0,589	0,151	0,351	0,293	OK
PAR_AP11	-1,528	0,182	0,807	0,306	OK
PAR_AP12	-1,594	0,185	0,817	-0,009	Pathological
PAR_AP13	0,119	0,104	0,457	0,127	Poor
PAR_AP14	0,478	0,149	0,376	0,435	OK
PAR_AP15	-1,066	0,162	0,728	0,199	OK
PAR_AP16	-0,637	0,150	0,639	0,391	OK
PAR_AP17	1,216	0,171	0,228	0,111	Poor
PAR_AP18	-2,056	0,154	0,874	0,128	Poor
PAR_AP19	-0,38	0,105	0,578	0,243	OK
PAR_AP20	-0,774	0,153	0,668	0,164	Poor
PAR_AP21	-0,442	0,147	0,594	0,249	OK
PAR_AP22	-0,571	0,149	0,624	0,186	Poor
PAR_AP23	0,844	0,113	0,296	0,075	Poor
PAR_AP24	0,733	0,111	0,319	0,207	OK
PAR_AP25	-0,103	0,104	0,513	0,217	OK
PAR_AP26	-0,211	0,145	0,540	0,131	Poor
PAR_AP27	-0,944	0,113	0,702	0,254	OK
PAR_AP28	-0,087	0,144	0,510	0,175	Poor
PAR_AP29	-1,7	0,191	0,832	0,270	OK
PAR_AP30	0,822	0,157	0,302	0,202	OK
PAR_AP31	1,026	0,117	0,262	0,132	Poor
PAR_AP32	-0,916	0,157	0,698	0,304	OK
PAR_AP33	-0,87	0,111	0,686	0,207	OK
PAR_AP34	-0,093	0,103	0,509	0,136	Poor
PAR_AP35	-0,315	0,146	0,564	0,167	Poor
PAR_AP36	-0,42	0,147	0,589	0,241	OK
PAR_AP37	0,435	0,148	0,386	0,428	OK
PAR_AP38	-0,715	0,109	0,653	0,233	OK
PAR_AP39	0,456	0,148	0,381	0,223	OK
PAR_AP40	-0,348	0,104	0,570	0,340	OK
PAR_AP41	0,269	0,105	0,423	0,369	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
PAR_AP42	0,055	0,104	0,474	0,222	OK
PAR_AP43	-1,312	0,171	0,772	0,241	OK
PAR_AP44	0,413	0,148	0,391	0,067	Poor
PAR_AP45	-1,1	0,116	0,732	0,292	OK
PAR_AP46	-1,839	0,143	0,849	0,180	Poor
PAR_AP47	-1,227	0,168	0,757	0,256	OK
PAR_AP48	-1,342	0,173	0,777	0,006	Poor
PAR_AP49	-1,021	0,114	0,717	0,301	OK
PAR_AP50	-1,978	0,210	0,866	0,066	Poor

Table A.10B Item parameters of *PCP-Parvularia* (Omitted the linking items)

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
PAR_BP01	-1,163	0,168	0,741	0,296	OK
PAR_BP02	-0,012	0,148	0,487	0,244	OK
PAR_BP03	-0,185	0,149	0,528	0,293	OK
PAR_BP05	-1,338	0,175	0,772	0,274	OK
PAR_BP07	-0,563	0,152	0,617	0,213	OK
PAR_BP08	0,642	0,156	0,337	0,155	Poor
PAR_BD12	-1,599	0,189	0,813	0,189	Poor
PAR_BP12	0,34	0,151	0,404	0,352	OK
PAR_BP13	-1,053	0,164	0,720	0,212	OK
PAR_BP14	-1,278	0,173	0,762	0,355	OK
PAR_BP15	0,715	0,158	0,321	0,232	OK
PAR_BP16	-0,339	0,150	0,565	0,200	OK
PAR_BP17	1,174	0,174	0,233	-0,053	Pathological
PAR_BP20	0,119	0,149	0,456	0,182	Poor
PAR_BP21	-1,708	0,195	0,829	0,181	Poor
PAR_BP22	-0,54	0,152	0,611	0,061	Poor
PAR_BP27	-0,702	0,155	0,648	0,277	OK
PAR_BP28	-0,923	0,160	0,694	0,060	Poor
PAR_BP29	-1,997	0,214	0,865	0,356	OK
PAR_BP30	1,33	0,181	0,207	0,018	Poor
PAR_BP31	-1,307	0,174	0,767	0,371	OK
PAR_BP33	-0,726	0,155	0,653	0,223	OK
PAR_BP36	0,119	0,149	0,456	0,368	OK
PAR_BP37	-0,974	0,162	0,705	0,181	Poor
PAR_BP38	-0,229	0,149	0,539	0,227	OK
PAR_BP45	-2,043	0,218	0,870	0,352	OK
PAR_BP46	-1,4	0,178	0,782	0,227	OK
PAR_BP47	-0,339	0,150	0,565	0,467	OK
PAR_BP48	-0,098	0,148	0,508	0,155	Poor
PAR_BP50	-0,339	0,150	0,565	0,208	OK

Table A.11A Item parameters of PCP-Media

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
MED_AP01	-0,416	0,109	0,638	0,198	OK
MED_AP02	-1,819	0,110	0,871	0,333	OK
MED_AP03	-0,358	0,108	0,625	0,148	Poor
MED_AP04	0,332	0,075	0,460	0,236	OK
MED_AP05	-0,034	0,075	0,548	0,166	Poor
MED_AP06	-0,144	0,106	0,576	0,025	Poor
MED_AP07	-0,733	0,082	0,703	0,273	OK
MED_AP08	-0,655	0,081	0,687	0,072	Poor
MED_AP09	-0,828	0,117	0,724	0,182	Poor
MED_AP10	-0,548	0,111	0,667	0,229	OK
MED_AP11	-1,381	0,096	0,816	0,300	OK
MED_AP12	-1,294	0,131	0,805	0,143	Poor
MED_AP13	-2,702	0,157	0,942	0,194	OK
MED_AP14	-0,001	0,105	0,542	0,465	OK
MED_AP15	-0,068	0,075	0,556	0,256	OK
MED_AP16	-0,349	0,077	0,621	0,116	Poor
MED_AP17	-0,897	0,119	0,737	0,305	OK
MED_AP18	-0,56	0,079	0,667	0,236	OK
MED_AP19	1,538	0,127	0,211	0,200	OK
MED_AP20	-2,264	0,184	0,914	0,277	OK
MED_AP21	0,478	0,106	0,427	0,292	OK
MED_AP22	-1,336	0,095	0,809	0,156	Poor
MED_AP23	-0,244	0,107	0,599	0,249	OK
MED_AP24	-1,759	0,108	0,865	0,310	OK
MED_AP25	-1,329	0,132	0,810	0,163	Poor
MED_AP26	-2,048	0,169	0,896	0,121	Poor
MED_AP27	0,674	0,077	0,381	0,202	OK
MED_AP28	-1,593	0,144	0,846	0,158	Poor
MED_AP29	0,138	0,075	0,506	0,226	OK
MED_AP30	0,184	0,105	0,497	0,158	Poor
MED_AP31	-0,256	0,107	0,602	0,211	OK
MED_AP32	-0,244	0,107	0,599	0,278	OK
MED_AP33	-1,055	0,123	0,766	0,266	OK
MED_AP34	-0,21	0,076	0,589	0,202	OK
MED_AP35	-0,335	0,108	0,620	0,144	Poor
MED_AP36	-0,597	0,112	0,677	0,182	Poor
MED_AP37	0,314	0,105	0,466	0,286	OK
MED_AP38	-0,285	0,077	0,607	0,274	OK
MED_AP39	-0,012	0,105	0,544	0,318	OK
MED_AP40	-0,659	0,113	0,690	0,248	OK
MED_AP41	-1,055	0,123	0,766	0,127	Poor
MED_AP42	-0,672	0,113	0,693	0,261	OK

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
MED_AP43	-0,279	0,077	0,605	0,167	Poor
MED_AP44	-1,301	0,094	0,804	0,373	OK
MED_AP45	-0,548	0,111	0,667	0,296	OK
MED_AP46	-0,998	0,086	0,753	0,113	Poor
MED_AP47	-0,535	0,111	0,664	0,104	Poor
MED_AP48	-1,132	0,125	0,779	0,208	OK
MED_AP49	-0,133	0,106	0,573	0,243	OK
MED_AP50	0,173	0,105	0,500	0,169	Poor

Table A.11b Item parameters of *PCP-Media* (Omitted the linking items)

Name/ Abbreviation	item difficulty IRT (B)	Standard Error of B	item difficulty (p)	item discrimination (Rit)	Flag for Rit
MED_BP03	0,703	0,111	0,373	0,315	OK
MED_BP04	0,158	0,107	0,500	0,304	OK
MED_BP06	0,444	0,108	0,432	0,322	OK
MED_BP08	-0,548	0,113	0,662	0,235	OK
MED_BP09	-1,273	0,132	0,797	0,262	OK
MED_BP11	0,363	0,108	0,451	0,288	OK
MED_BP13	0,826	0,112	0,346	0,189	Poor
MED_BP16	0,17	0,107	0,497	0,502	OK
MED_BP18	-0,267	0,109	0,600	0,298	OK
MED_BP20	-0,291	0,110	0,605	0,149	Poor
MED_BP22	-1,189	0,129	0,784	0,255	OK
MED_BP23	-1,015	0,124	0,754	0,405	OK
MED_BP24	-0,911	0,121	0,735	0,341	OK
MED_BP26	-1,683	0,150	0,854	0,284	OK
MED_BP27	0,691	0,111	0,376	0,116	Poor
MED_BP28	-1,515	0,142	0,832	0,173	Poor
MED_BP31	-0,058	0,108	0,551	0,332	OK
MED_BP32	-1,979	0,166	0,886	0,227	OK
MED_BP33	-1,437	0,139	0,822	0,311	OK
MED_BP34	-1,475	0,140	0,827	0,350	OK
MED_BP35	-0,256	0,109	0,597	0,399	OK
MED_BP37	-0,047	0,108	0,549	0,161	Poor
MED_BP38	-0,664	0,115	0,686	0,173	Poor
MED_BP39	-1,344	0,135	0,808	0,309	OK
MED_BP41	-0,703	0,116	0,695	0,250	OK
MED_BP42	-2,559	0,209	0,932	0,268	OK
MED_BP44	-1,683	0,150	0,854	0,219	OK
MED_BP45	-2,559	0,209	0,932	0,180	Poor
MED_BP48	-0,387	0,111	0,627	0,206	OK
MED_BP49	-0,771	0,118	0,708	0,286	OK
MED_BP50	-0,498	0,112	0,651	0,312	OK

APPENDIX B

CHARACTERISTICS OF THE FLAGGED ITEMS IN *INICÍA*

Table B.1A: Poor or pathological items in *PCD-Básica*

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
1	0,92	0,03	0,00	ABC	There is no REAL alternative for the correct answer
2	0,86	0,15	0,12	A	There is no REAL alternative for the correct answer
4	0,84	0,19	0,16	A	There is no REAL alternative for the correct answer
5	0,59	0,15	0,10	BD	The weakest students find the correct alternative too easily
7	0,36	0,27	0,22	D	
12	0,77	0,19	0,14	A	There is no REAL alternative for the correct answer
17	0,84	0,15	0,11	A	There is no REAL alternative for the correct answer
26	0,44	0,13	0,08	A	The BEST students do not find the correct alternative and the weakest students find the correct alternative too easily
34	0,27	0,14	0,09	A	The BEST students do not find the correct alternative and the weakest students find the correct alternative too easily
43	0,36	-0,01	-0,06	ABCD	There seems to be several (or NO) correct answer. High guessing
45	0,85	0,16	0,12	A	There is no REAL alternative for the correct answer
47	0,53	0,16	0,11	A	There seems to be TWO correct answers (B and A).
50	0,26	0,15	0,10	A	There seems to be several (or NO) correct answer. The BEST students do not find the correct alternative and the weakest students find the correct alternative too easily High guessing.
55	0,84	0,14	0,10	A	There is no REAL alternative for the correct answer
56	0,73	0,09	0,04	ABD	There is no REAL alternative for the correct answer
57	0,33	0,17	0,12	A	There seems to be several (or NO) correct answer. The BEST students do not find the correct alternative
61	0,83	0,15	0,11	A	There is no REAL alternative for the correct answer
62	0,36	0,18	0,13	A	There seems to be TWO correct answers (C and A). The BEST students do not find the correct alternative
64	0,09	0,16	0,13	A	There is no REAL alternative for the correct answer. Check the key! Most BEST students selected the alternative D.
68	0,47	0,11	0,05	AB	There seems to be several (or NO) correct answer. The weakest students find the correct alternative too easily
75	0,14	0,05	0,01	ABD	There seems to be several (or NO) correct answer. The BEST students do not find the correct alternative and the weakest students find the correct alternative too easily
76	0,75	0,17	0,12	A	There is no REAL alternative for the correct answer

1) A: Rit < 0.20 item-total correlation is low, B: Rir >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high

Figure B.1A Poor or pathological items in PCD-Básica Version A

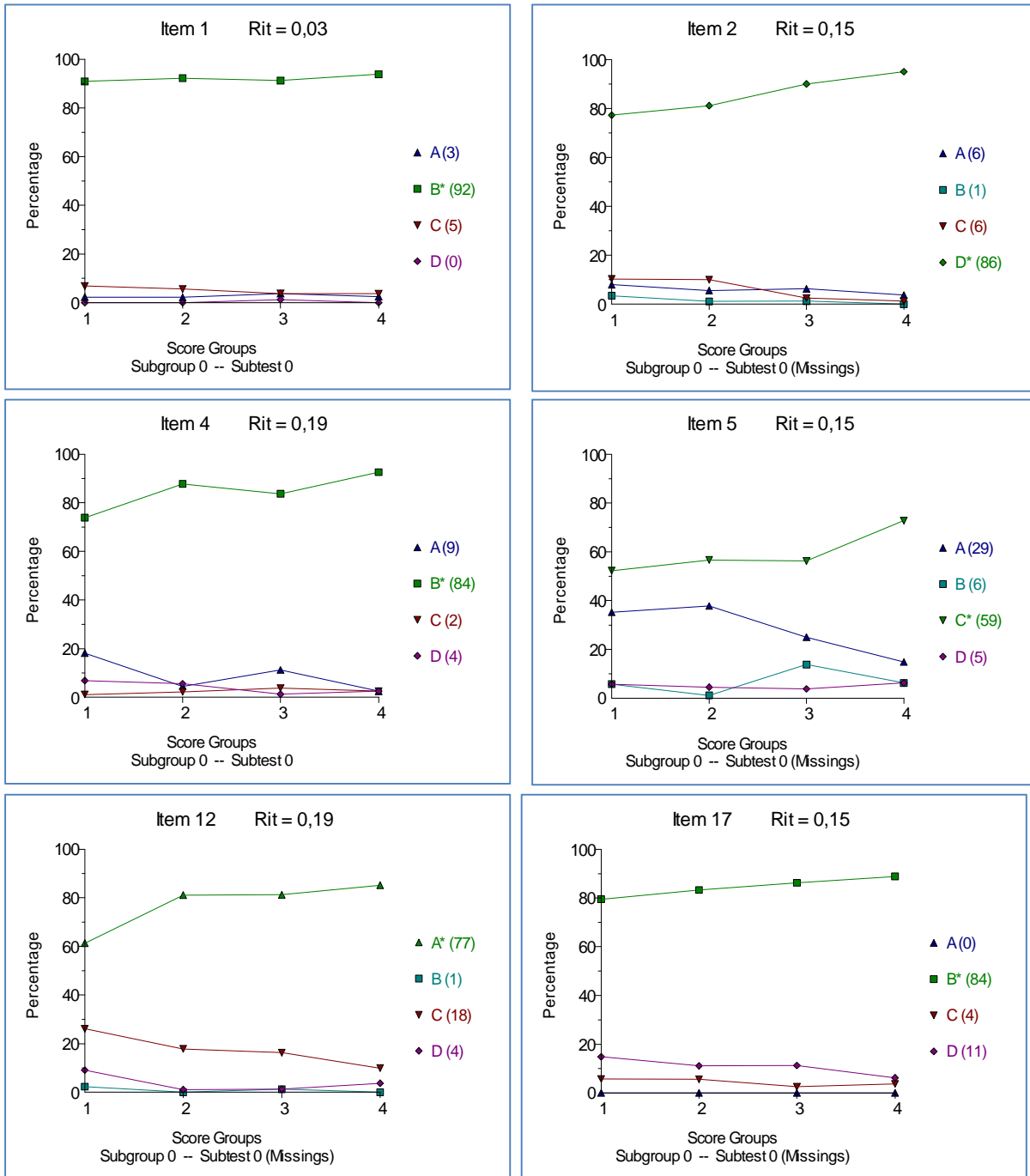


Figure B.1A Poor or pathological items in PCD-Básica Version A (cont'd.)

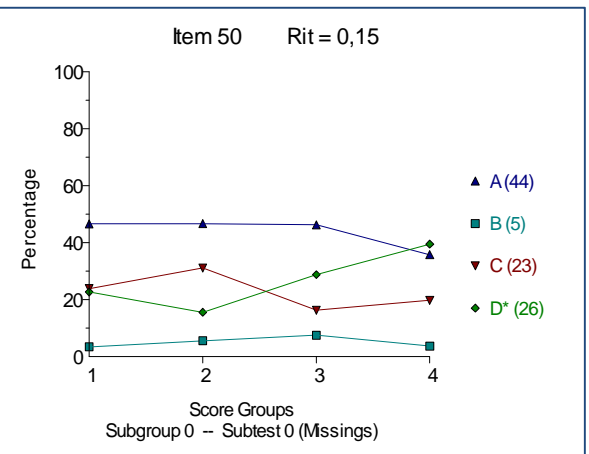
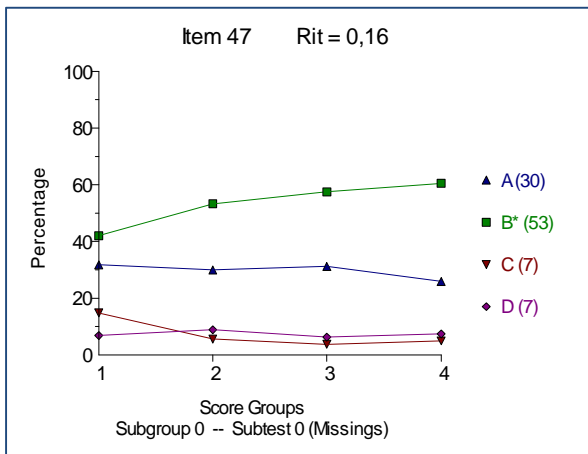
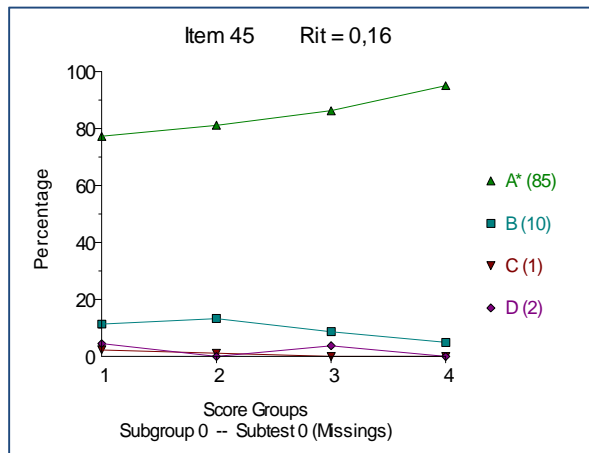
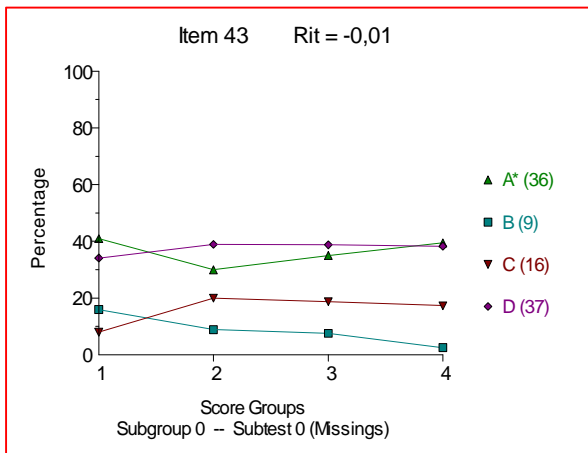
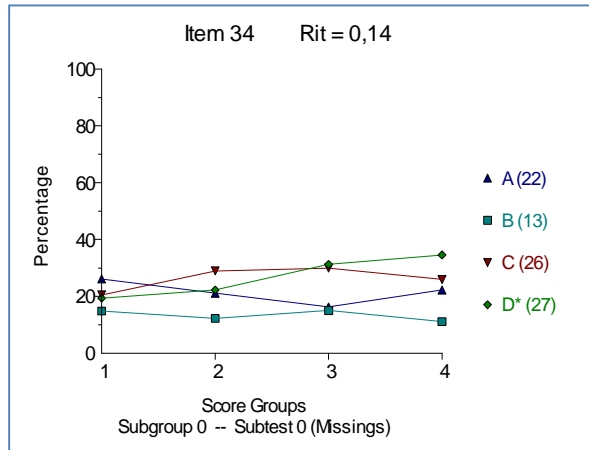
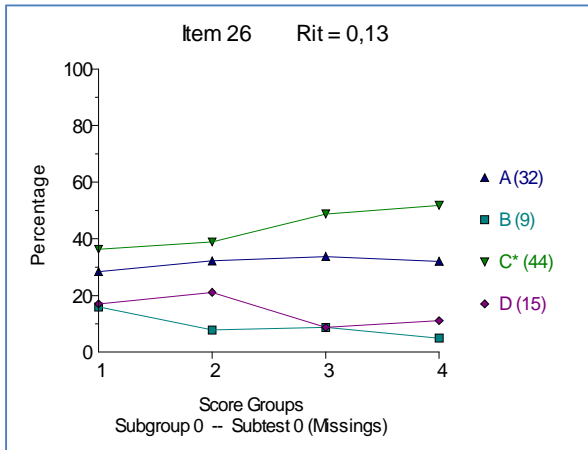


Figure B.1A Poor or pathological items in PCD-Básica Version A (cont'd.)

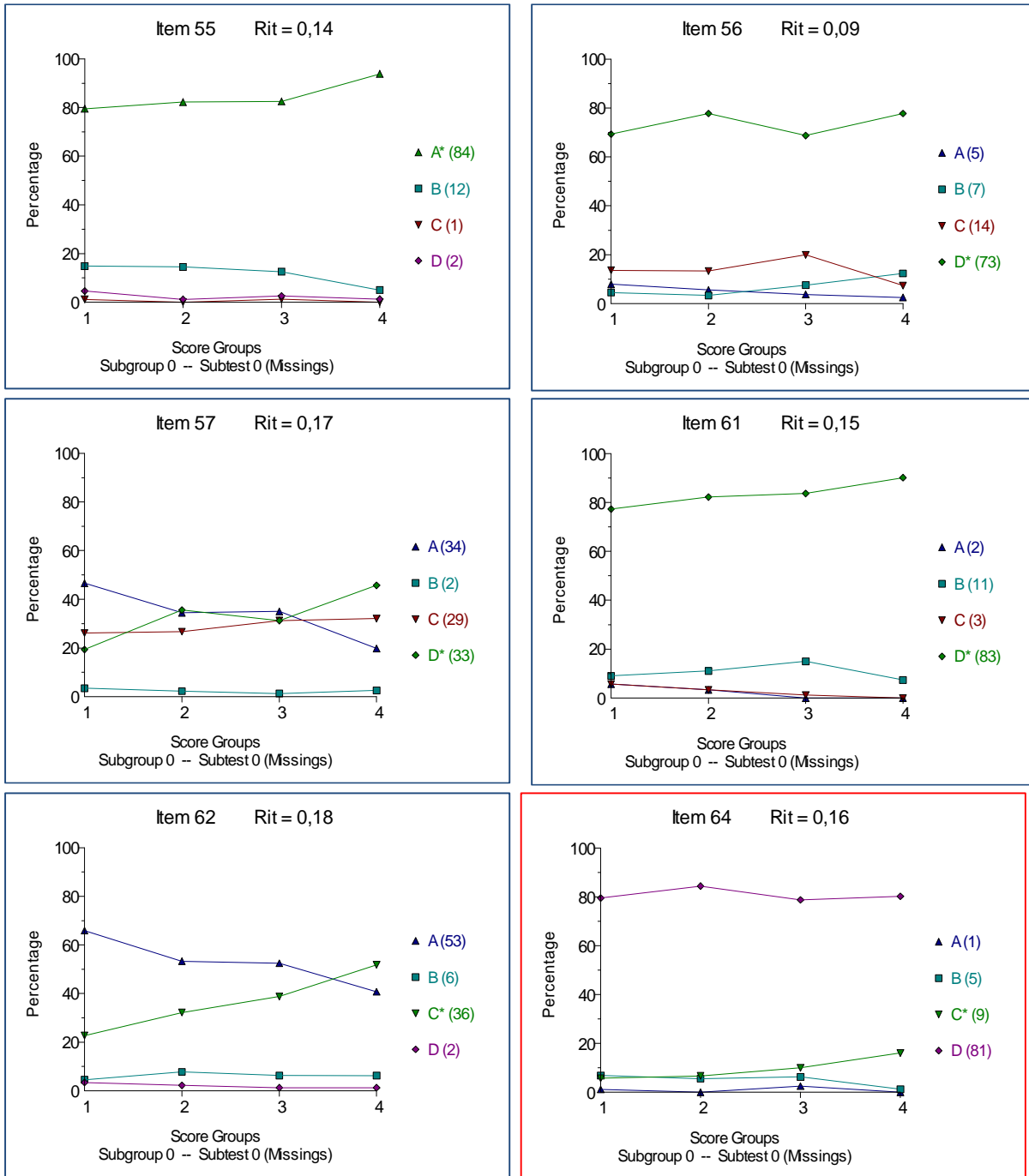


Figure B.1A Poor or pathological items in PCD-Básica Version A (cont'd.)

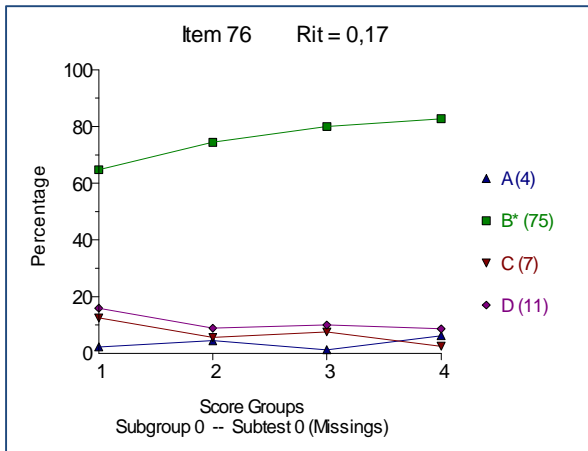
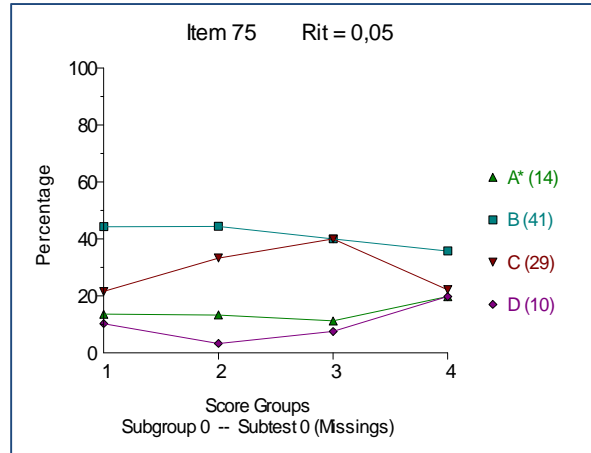
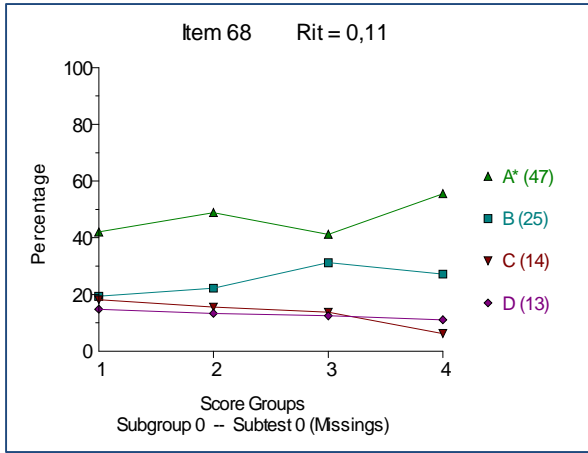


Table B.1B: Poor or pathological items in *PCD-Básica* Version B

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
1	0,9	0,05	0,02	A	There is no REAL alternative for the correct answer
2	0,9	0,07	0,04	A	There is no REAL alternative for the correct answer
9	0,48	0,11	0,06	AB	The weakest students find the correct alternative too easily
10	0,85	0,19	0,16	A	There is no REAL alternative for the correct answer
13	0,41	0,12	0,07	AB	The weakest students find the correct alternative too easily
17	0,69	0,17	0,12	A	There is no REAL alternative for the correct answer
20	0,35	0,18	0,13	BD	The BEST students do not find the correct alternative
34	0,32	0,14	0,09	ABD	There seems to be TWO correct answers (A and D)
39	0,31	0,18	0,13	A	The BEST students do not find the correct alternative
43	0,39	0,13	0,08	A	The BEST students do not find the correct alternative
45	0,57	0,19	0,14	A	The BEST students are messing with D
53	0,51	0,16	0,11	A	The BEST students are messing with B and D
55	0,64	0,16	0,11	A	The BEST students are messing with D
56	0,64	0,18	0,14	A	The BEST students are messing with D
57	0,39	0,22	0,18	D	
61	0,78	0,06	0,01	AB	There is no REAL alternative for the correct answer
62	0,33	0,16	0,11	A	There seems to be TWO correct answers (A and C)
64	0,39	0,12	0,07	ABD	There seems to be TWO correct answers (D and C)
71	0,46	0,26	0,21	D	
75	0,24	0,16	0,12	A	There seems to be several correct answers
76	0,41	0,16	0,11	A	The weakest students find the correct alternative too easily
77	0,59	0,13	0,08	A	There is no REAL alternative for the correct answer
80	0,87	0,11	0,08	A	There is no REAL alternative for the correct answer

1)A: Rit < 0.20 item-total correlation is low, B: Rir >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rir >= 10 a distracter - test score correlation is suspiciously high

Figure B.1B Poor or pathological items in PCD-Básica Version B

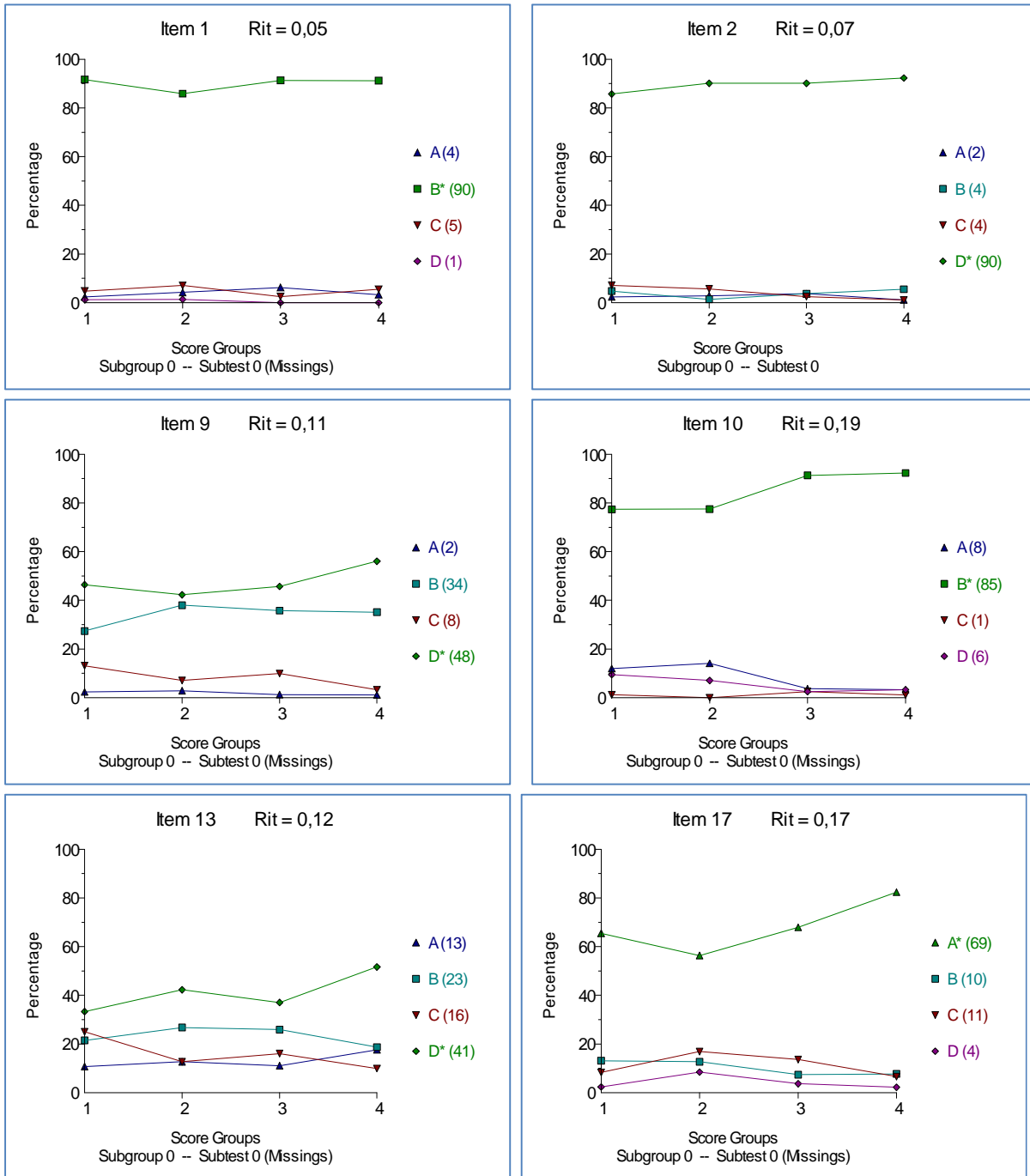


Figure B.1B Poor or pathological items in PCD-Básica Version B (cont'd.)

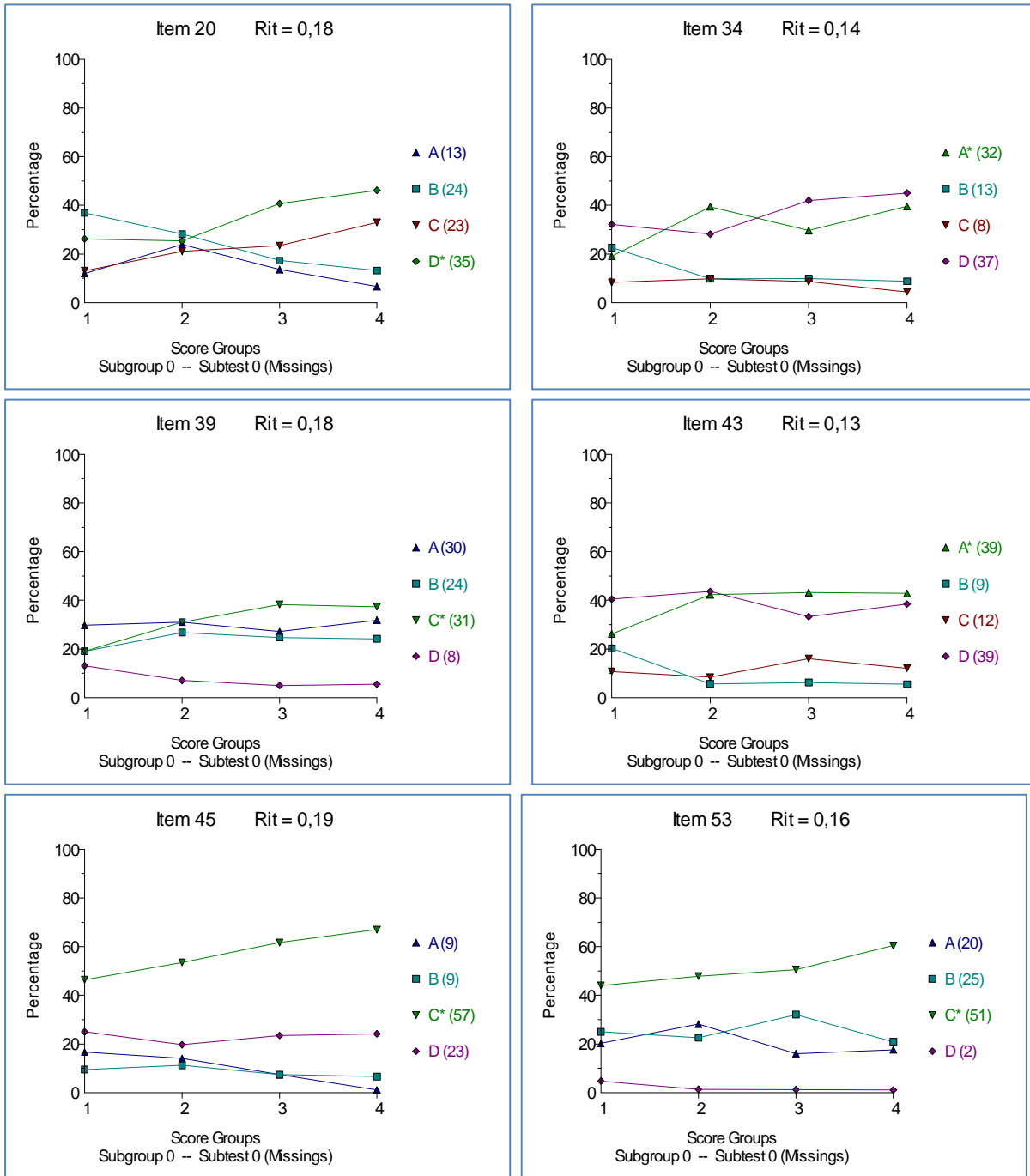


Figure B.1B Poor or pathological items in PCD-Básica Version B (cont'd.)

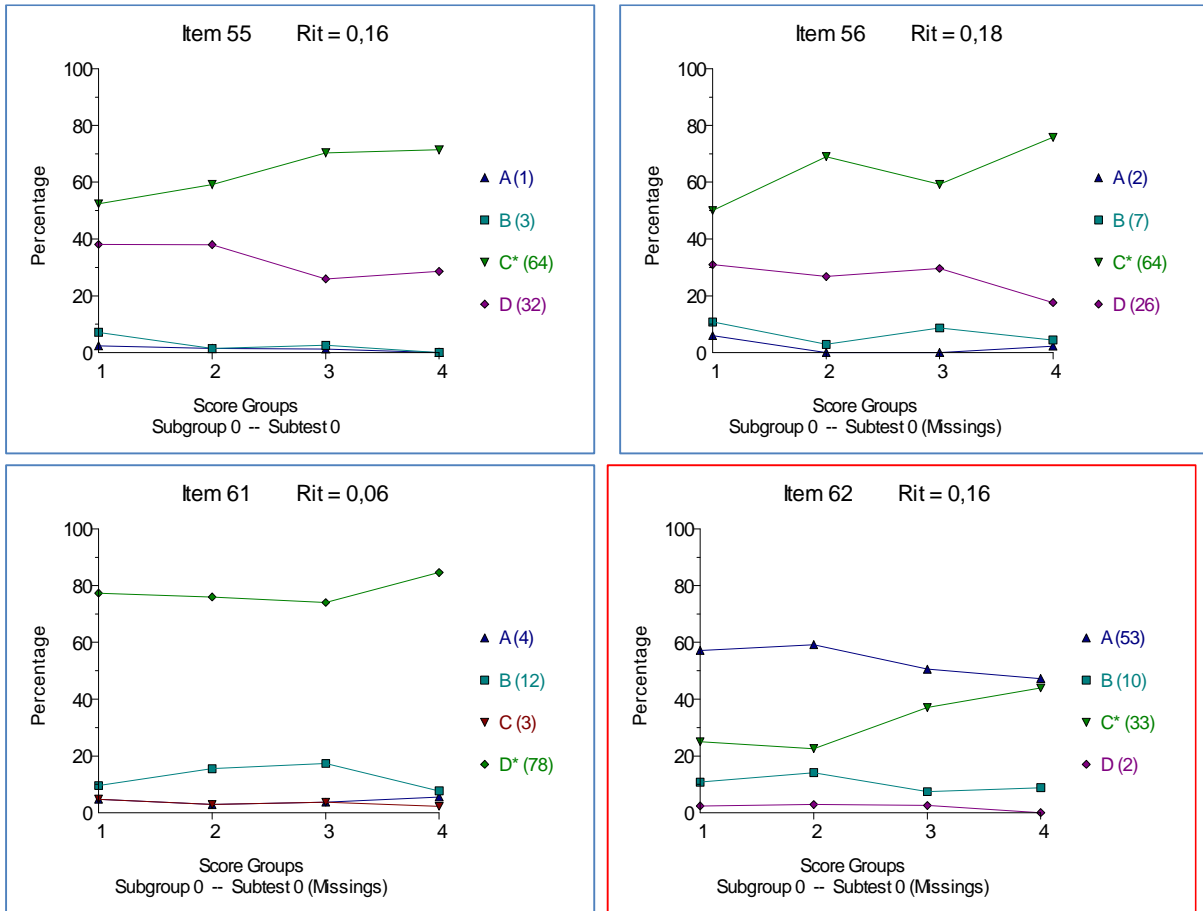


Figure B.1B Poor or pathological items in PCD-Básica Version B (cont'd.)

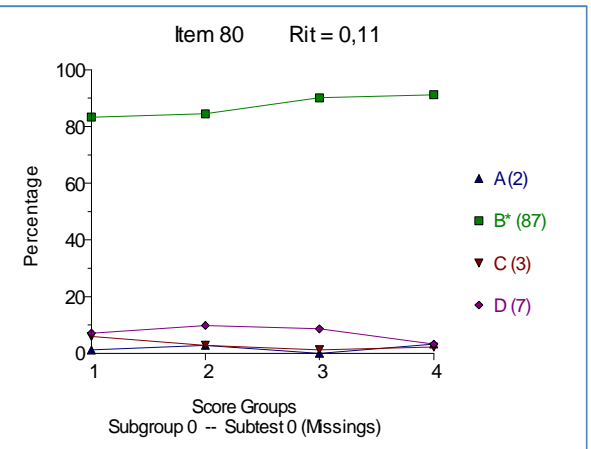
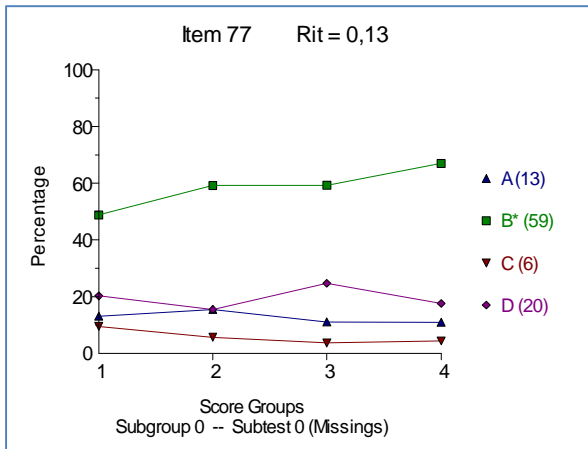
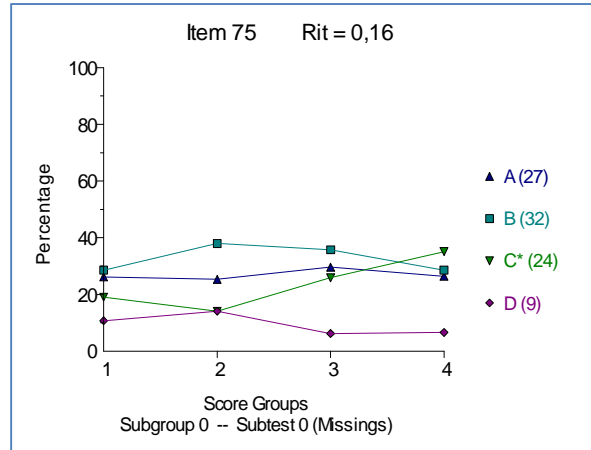
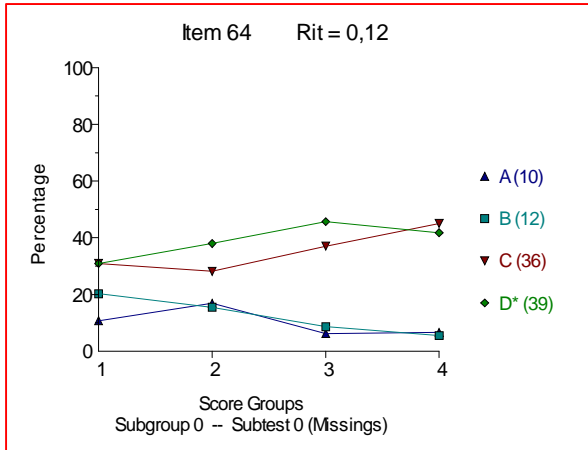
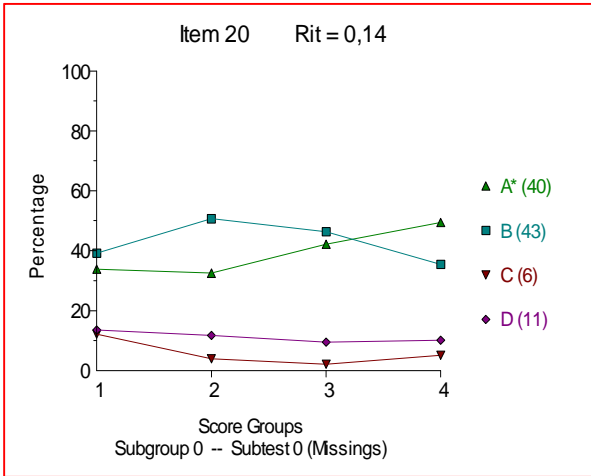
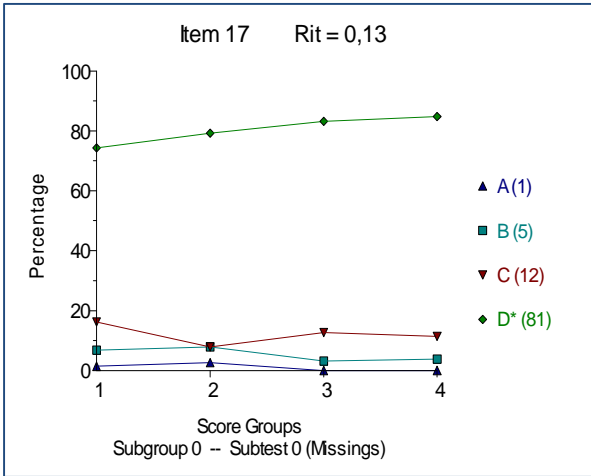
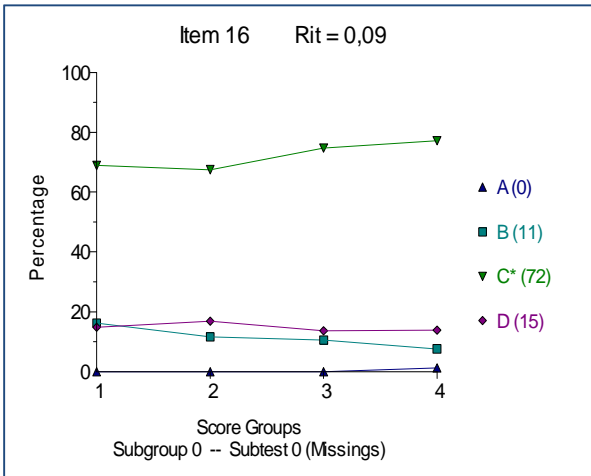
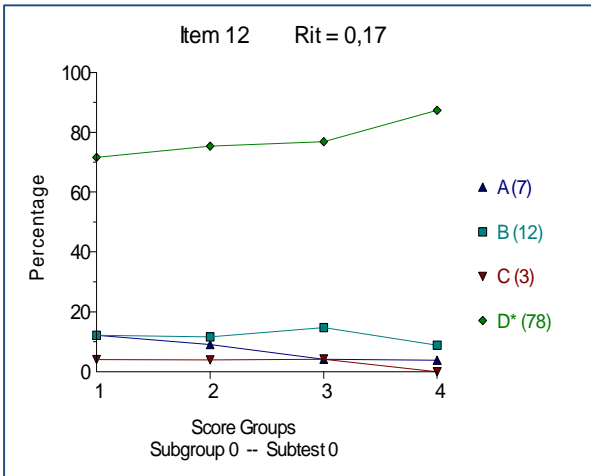
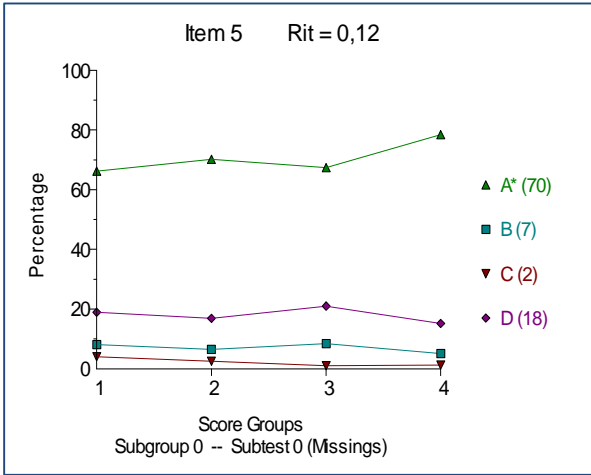
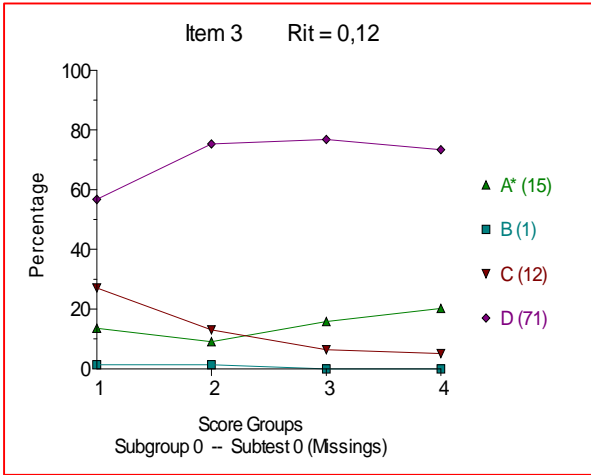


Table B.2A Poor or pathological items in PCP-Básica Version A

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
3	0,15	0,12	0,05	ABD	The BEST ones do not find the correct answer but they are distracted by D. Check the key. Is D the real key?
5	0,7	0,12	0,04	A	There is no REAL alternatives for the correct answer and the WEAKEST ones find the correct answer too easily
12	0,78	0,17	0,09	A	There is no REAL alternative for the correct answer
16	0,72	0,09	0,01	AB	There is no REAL alternative for the correct answer and the weakest students find the correct alternative too easily
17	0,81	0,13	0,06	A	There is no REAL alternative for the correct answer and the weakest students find the correct alternative too easily
20	0,4	0,14	0,05	A	There seems to be TWO correct answers (A and B)
22	0,63	0,18	0,09	A	The WEAKEST ones find the correct alternative too easily.
34	0,74	0,18	0,1	A	There is no REAL alternative for the correct answer and the weakest students find the correct alternative too easily
39	0,36	0,12	0,04	AB	There seems to be TWO alternatives for the correct answer (C and B)
41	0,9	0,09	0,03	A	There is no REAL alternative for the correct answer and the weakest students find the correct alternative too easily
47	0,77	0,14	0,07	A	There is no REAL alternative for the correct answer and the weakest students find the correct alternative too easily
49	0,52	0,16	0,07	AB	There seems to be TWO alternatives for the correct answer (D and B)

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high



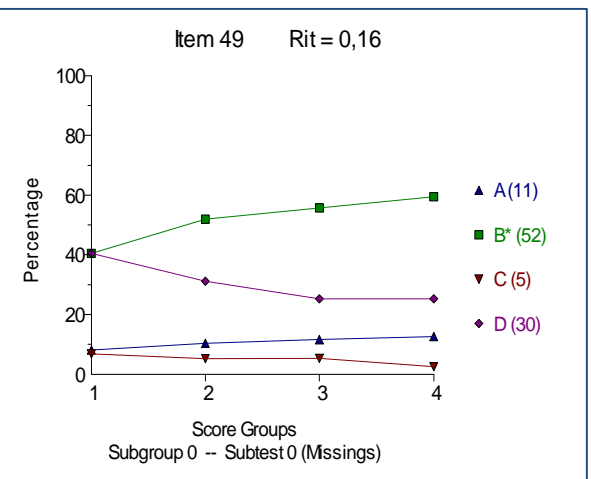
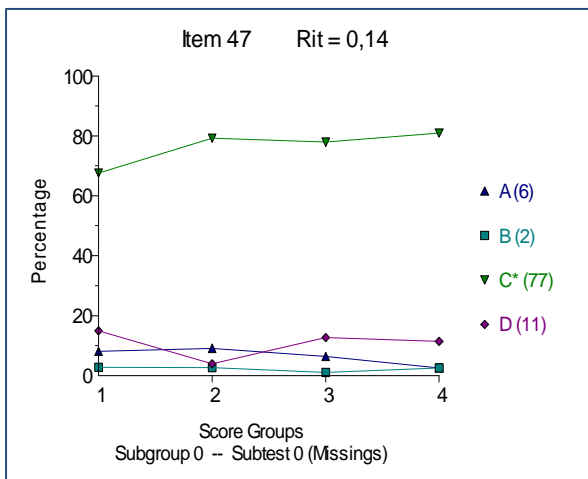
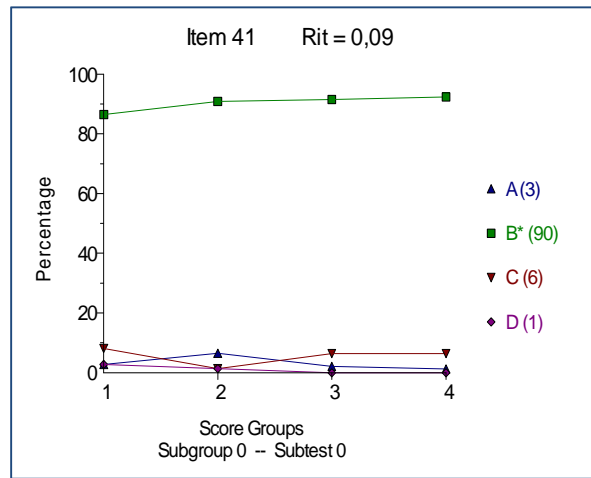
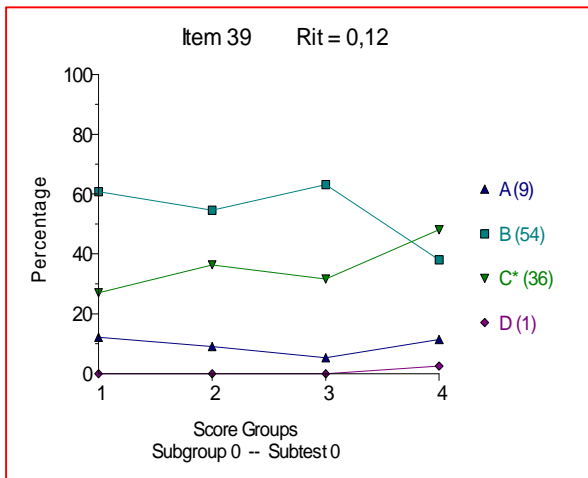
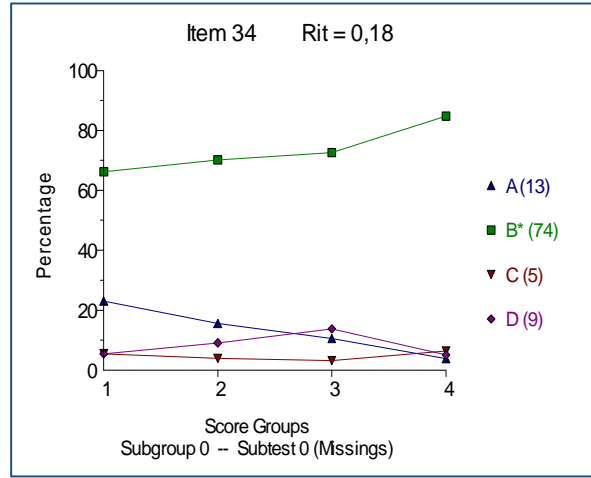
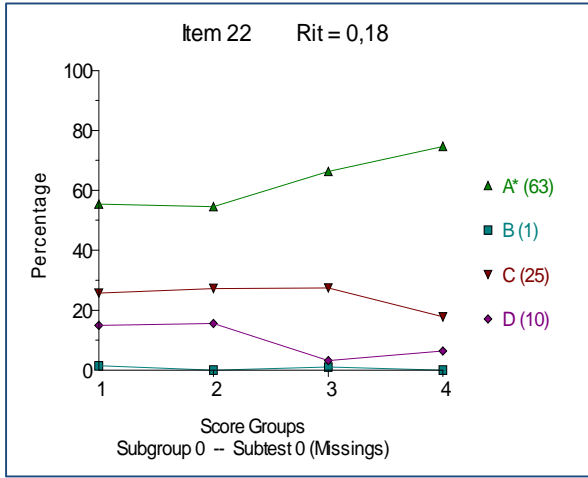
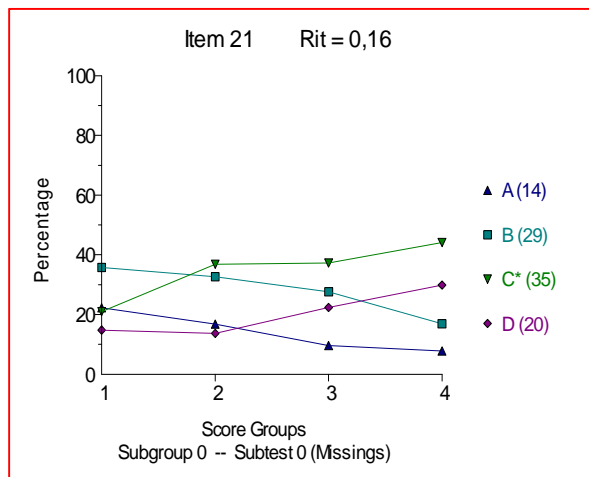
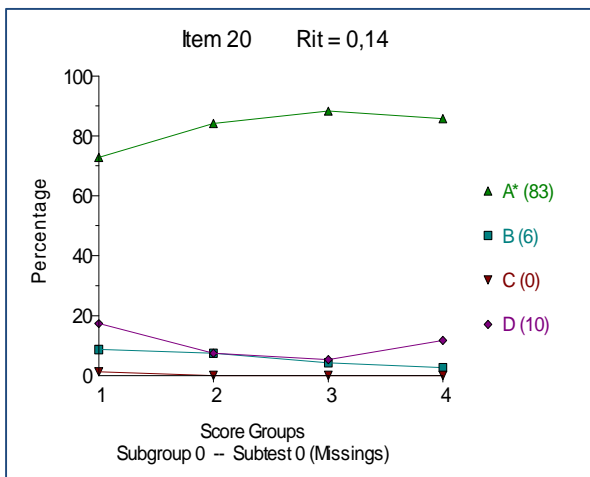
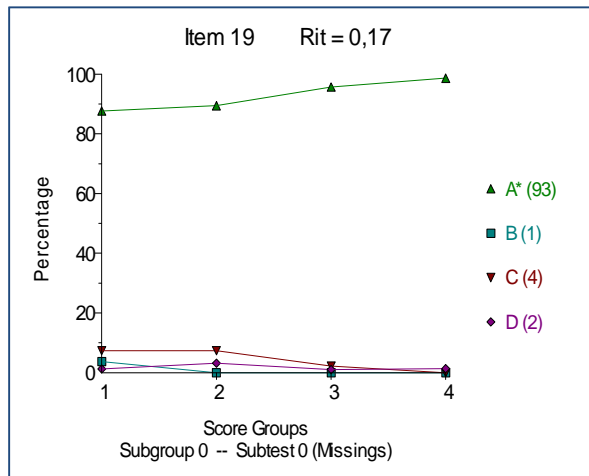
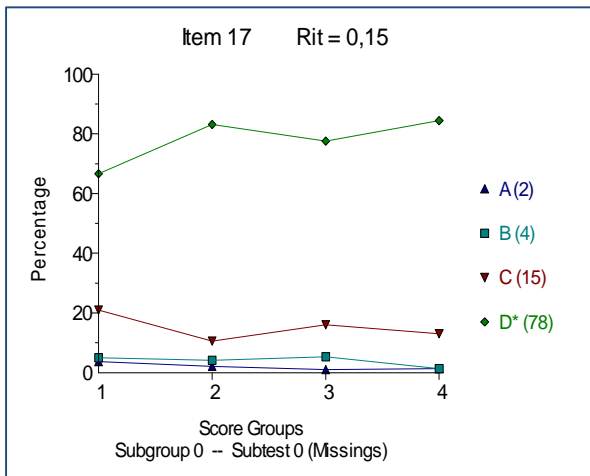
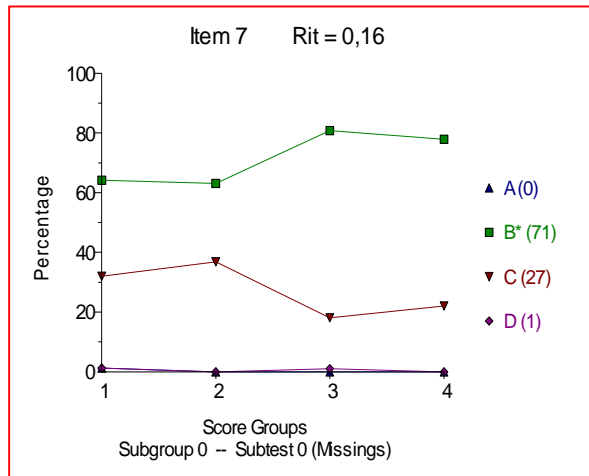
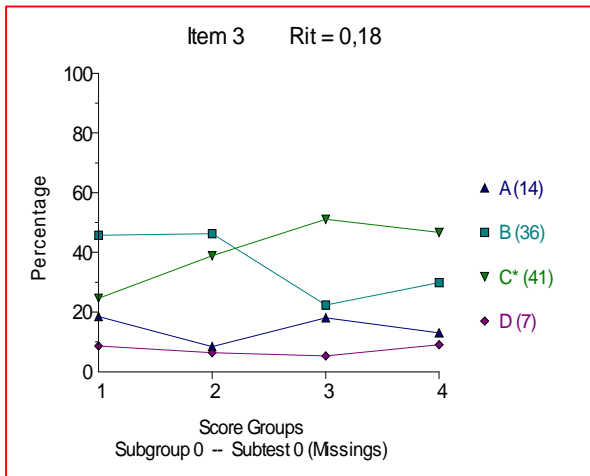


Table B.2B Poor or pathological items in PCP-Básica Version B

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
3	0,41	0,18	0,09	A	There seems to be TWO alternatives for the correct answer (C and B) and the BEST ones are distracted by B
7	0,71	0,16	0,08	A	There seems to be TWO alternatives for the correct answer (C and B) and the BEST ones are distracted by C
17	0,78	0,15	0,07	A	There is no REAL alternatives for the correct answer
19	0,93	0,17	0,12	A	There is no REAL alternatives for the correct answer
20	0,83	0,14	0,07	A	The BEST ones are distracted by D
21	0,35	0,16	0,07	ABD	There is NO correct answer and the BEST ones are distracted by D. Check the key!
23	0,69	0,1	0,01	A	The BEST ones are distracted by C
31	0,8	0,08	0	AB	The BEST ones are distracted by D and the POOREST ones find the correct answer too easily
33	0,64	0,12	0,03	AB	The BEST ones are distracted by B and the POOREST ones find the correct answer too easily
37	0,76	0,18	0,1	A	no problem
42	0,84	0,16	0,09	A	There is no REAL alternative for the correct answer and the POOREST students find the correct alternative too easily
47	0,71	0,19	0,11	A	no problem
48	0,88	0,09	0,03	AB	There is no REAL alternatives for the correct answer
50	0,32	0,19	0,1	A	The BEST students are messing with A and D. There seems to be NO correct answer

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high



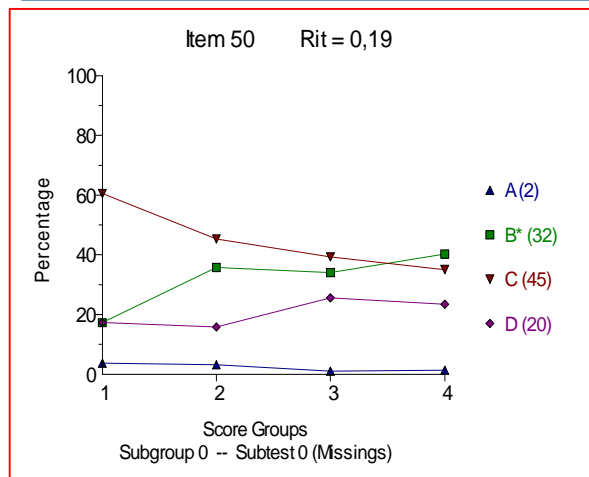
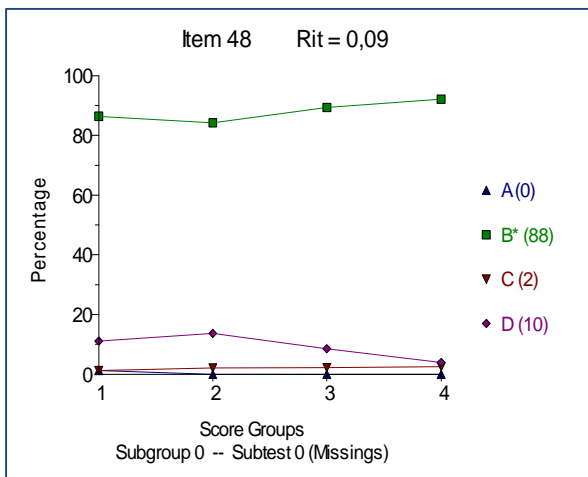
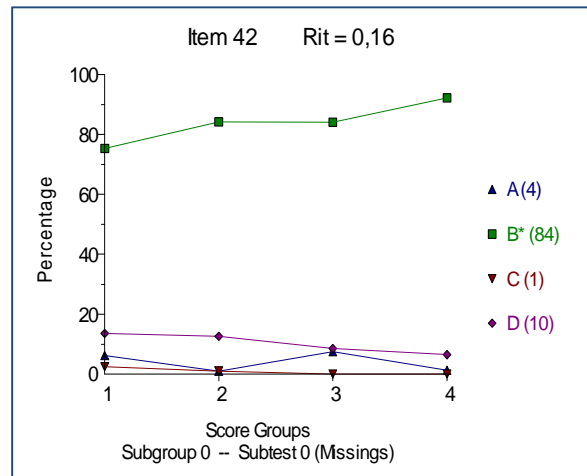
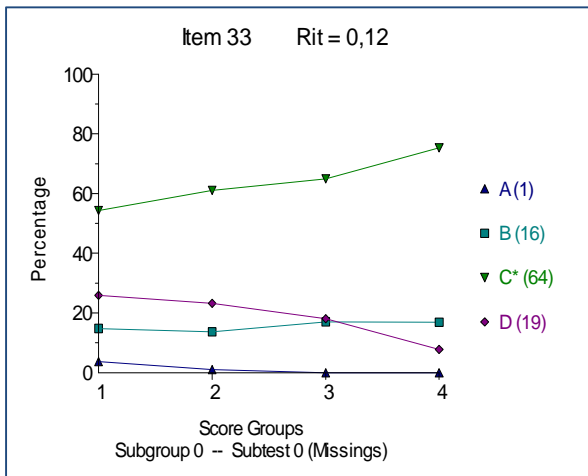
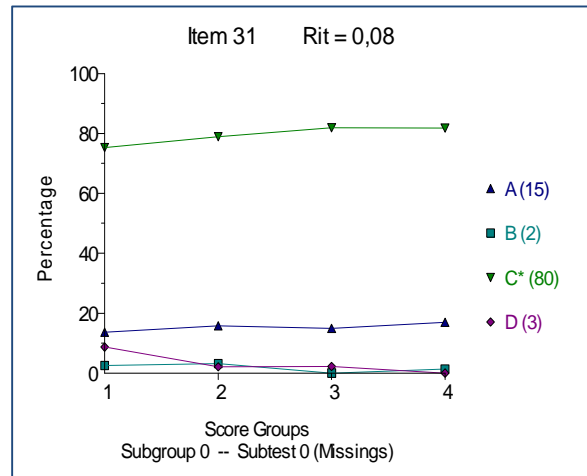
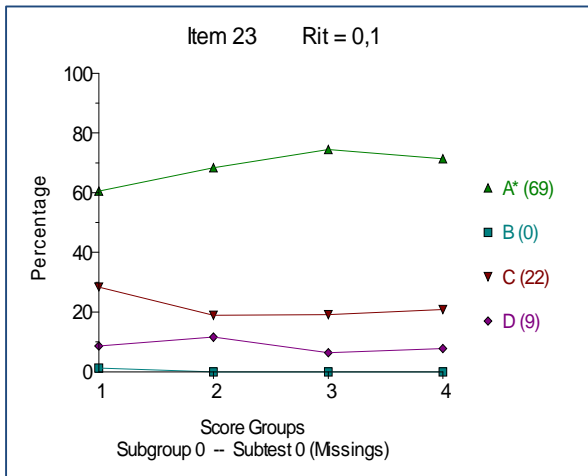
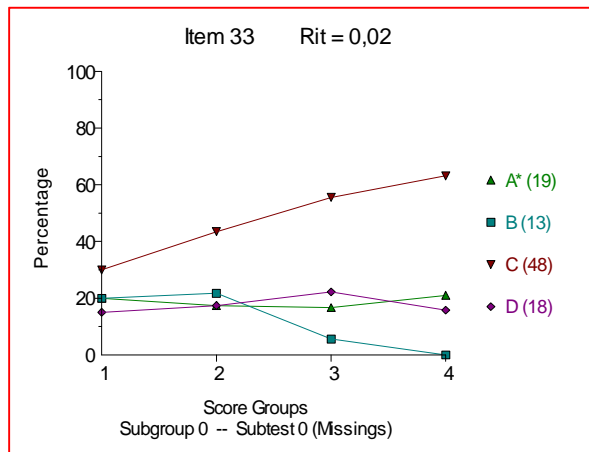
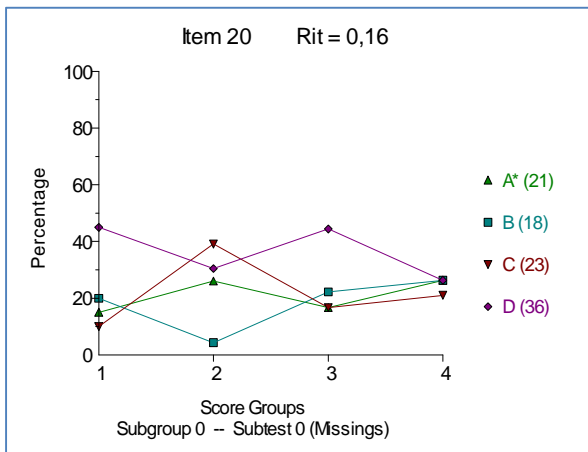
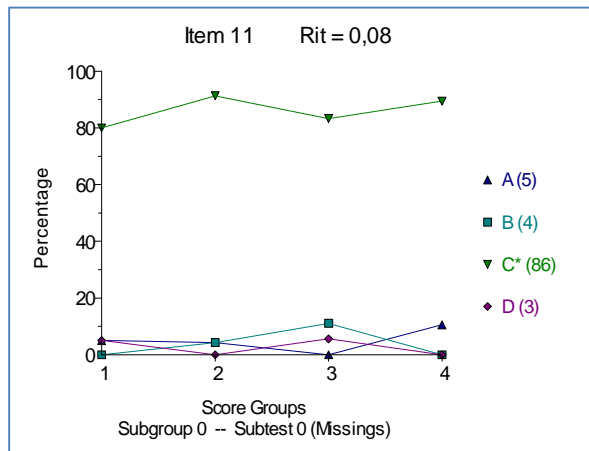
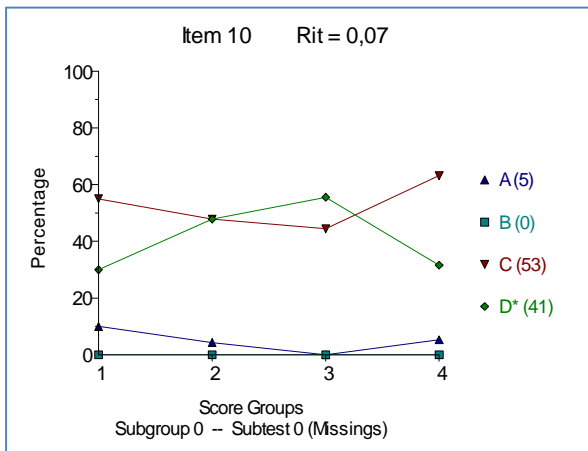
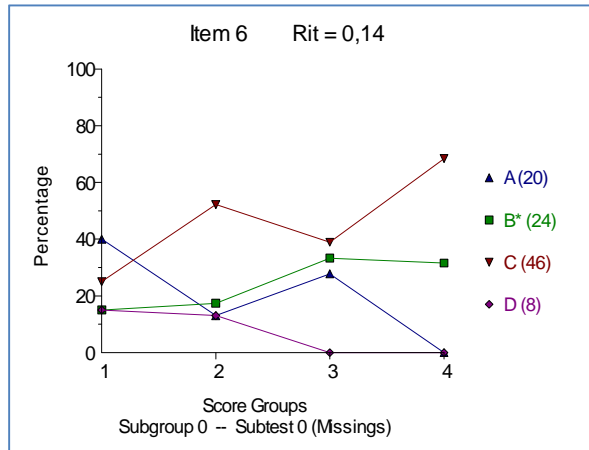
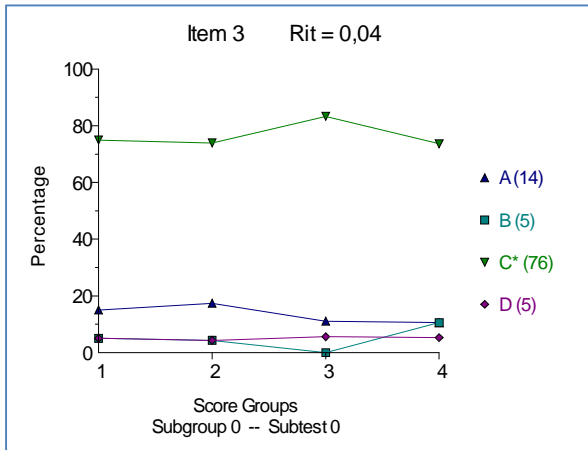
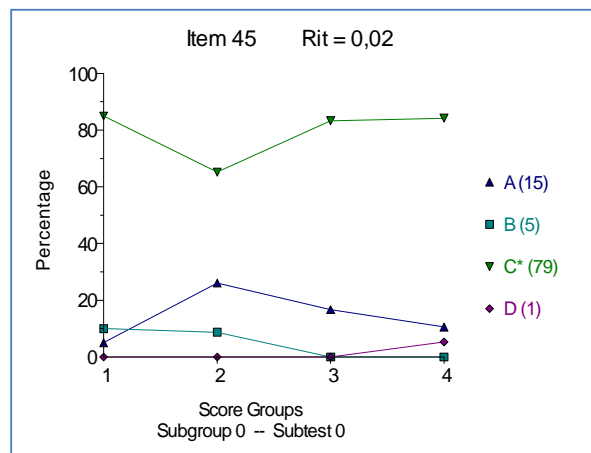
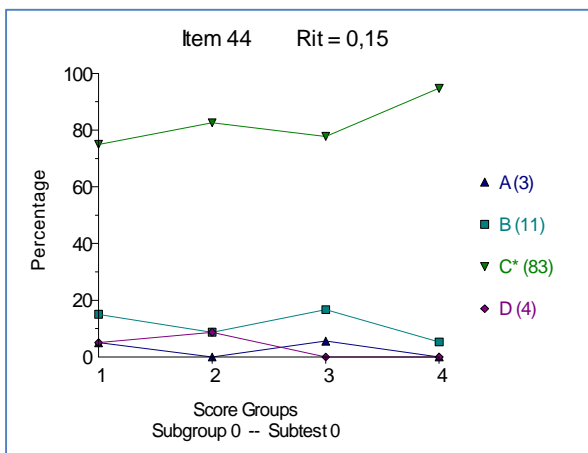
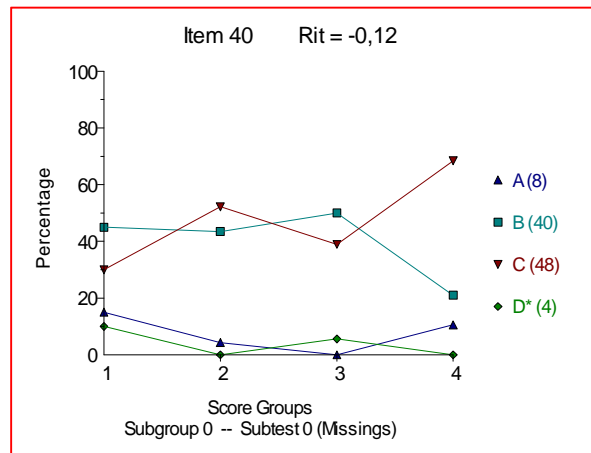
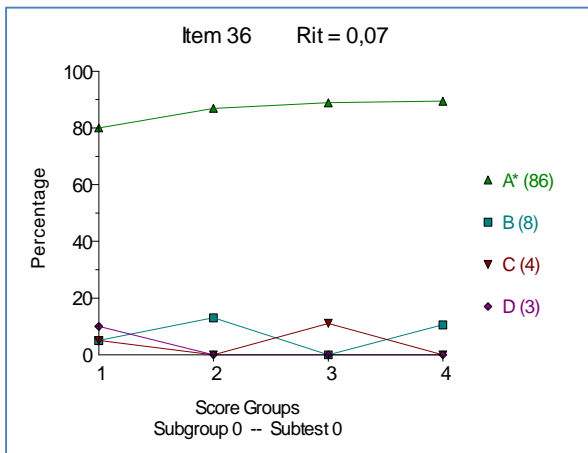
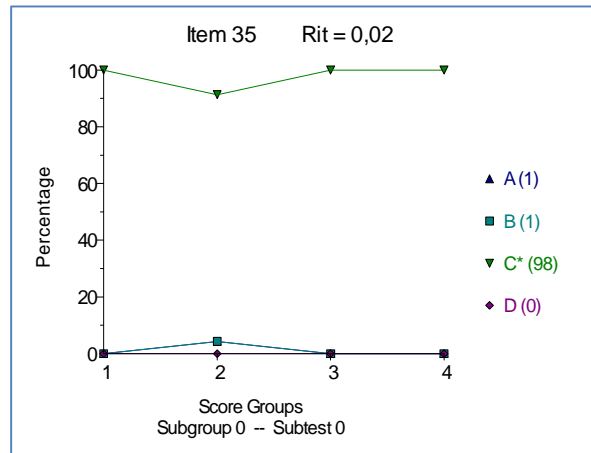
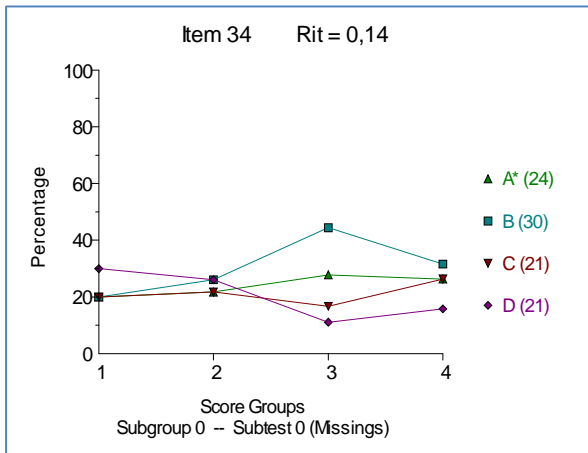


Table B.3 Poor or pathological items in PCD-Biología

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
2	0,68	0,29	0,23	D	
3	0,76	0,04	-0,02	ABC	There is no REAL alternative for the correct answer
4	0,76	0,25	0,19	D	
6	0,24	0,14	0,08	ABD	The BEST students are distracted to alternative C (Check the key!)
10	0,41	0,07	0	AB	The BEST students are distracted to alternative C (Check the key!)
11	0,86	0,08	0,03	AB	There is no REAL alternative for the correct answer
12	0,74	0,25	0,19	D	
16	0,64	0,24	0,17	D	
18	0,26	0,19	0,13	A	no problem
20	0,21	0,16	0,1	A	There is no REAL correct answer
28	0,8	0,35	0,3	D	
33	0,19	0,02	-0,03	ABCD	This is pathological item. The real correct answer seems to be C (not A). Check the key!
34	0,24	0,14	0,08	BD	There is no REAL correct answer
35	0,97	0,02	0	AB	There is no REAL alternative for the correct answer
36	0,86	0,07	0,02	A	There is no REAL alternative for the correct answer
40	0,04	-0,12	-0,15	ABCD	This is pathological one because the BEST students are distracted to alternative C (Check the key!). Definitely D is not the Key! I'd guess C instead.
44	0,82	0,15	0,09	A	There is no REAL alternative for the correct answer
45	0,79	0,02	-0,04	ABCD	There is no REAL alternative for the correct answer
47	0,04	-0,12	-0,15	ABCD	This is pathological one because the BEST students are distracted to alternative D (Check the key!). Definitely A is not the Key! I'd guess D instead.
50	0,61	0,11	0,04	AB	There is no REAL alternative for the correct answer
52	0,49	-0,11	-0,18	ABCD	This is pathological one because the BEST students are confused. For the best students there are two correct answers (C and B)
53	0,13	0,12	0,07	A	There is no REAL alternative for the correct answer
55	0,64	0,12	0,05	ABD	There is no REAL alternative for the correct answer
58	0,49	0,08	0,01	ABD	The BEST students are confused. For the best students there are TWO correct answers (B and D)

1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high





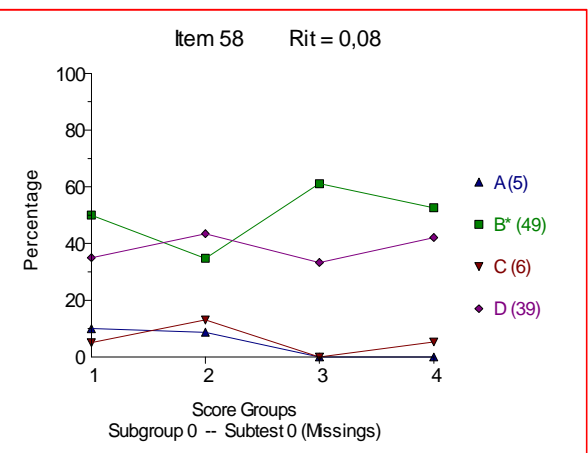
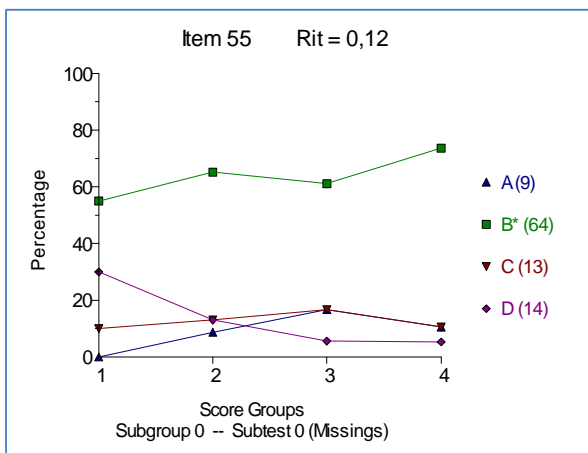
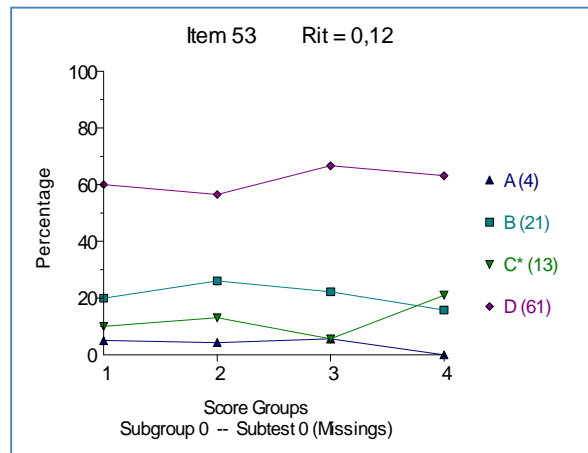
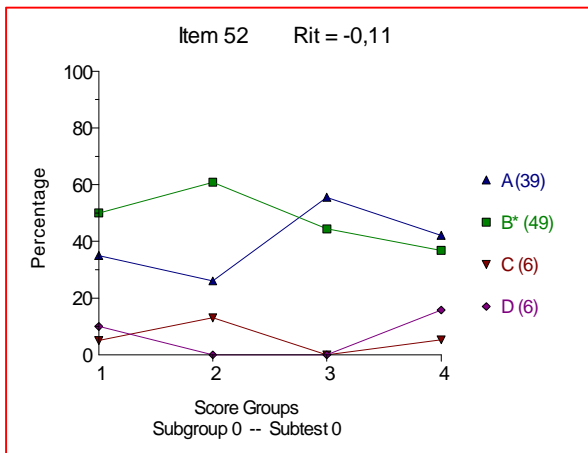
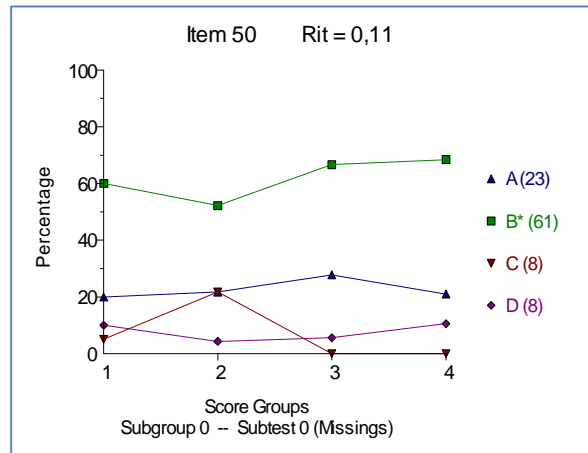
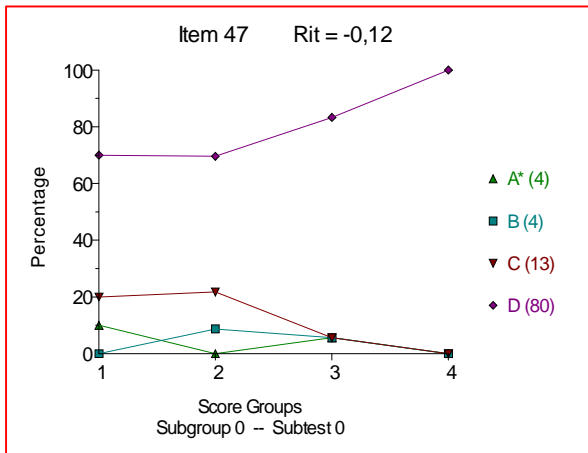


Table B.4 Poor or pathological items in PCD-Física

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
10	0,48	0,47	0,44	D	
16	0,43	0,44	0,41	D	
24	0,22	0,30	0,27	D	
25	0,57	0,31	0,27	D	
30	0,35	0,29	0,25	D	
33	0,29	0,52	0,49	D	
34	0,37	0,52	0,49	D	
39	0,49	0,22	0,18	D	
40	0,2	0,34	0,31	D	
43	0,89	0,12	0,09	A	There is no REAL alternative for the correct answer
44	0,58	0,16	0,12	A	The BEST students are distracted to alternative B. For the best ones there are TWO correct answers (A and C).
46	0,51	0,17	0,13	A	The POOREST students guess the correct answer
53	0,54	0,25	0,21	D	
56	0,88	0,14	0,11	A	There is no REAL alternative for the correct answer
57	0,55	0,36	0,32	D	
59	0,32	0,13	0,09	ABD	There seems to be TWO alternatives for the correct answer; the BEST students are distracted to alternative B (Check the key!)

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high

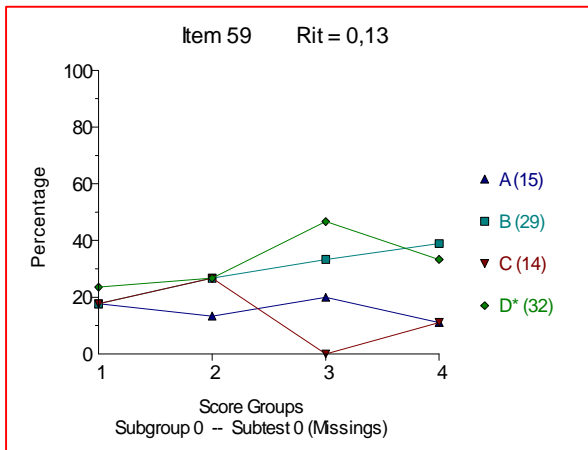
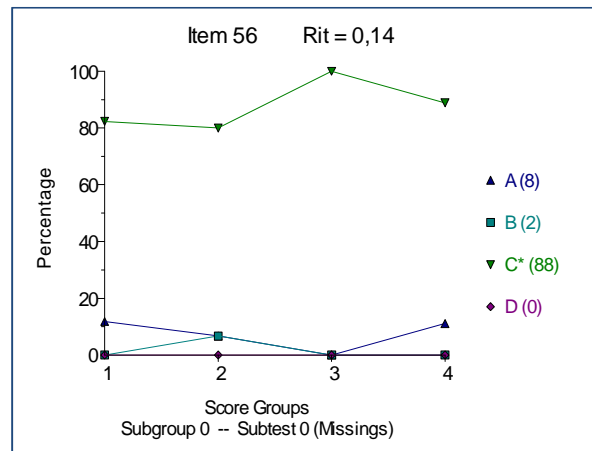
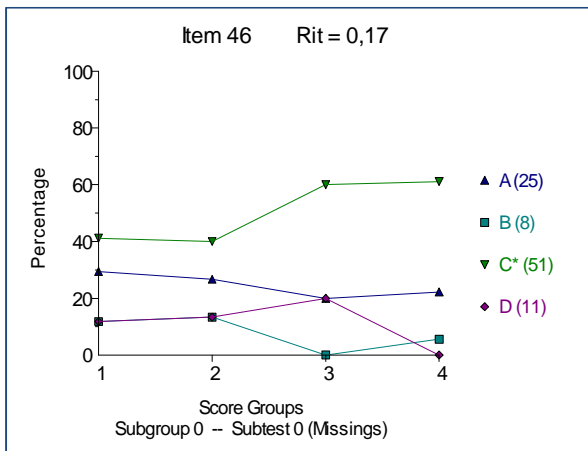
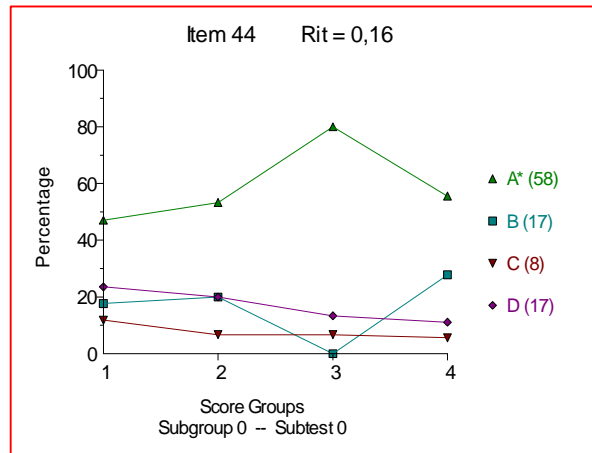
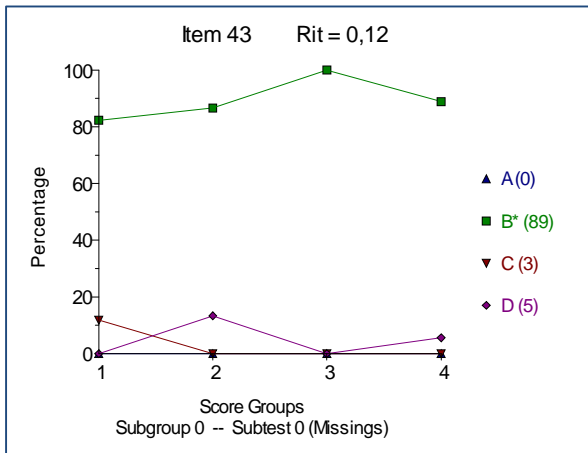
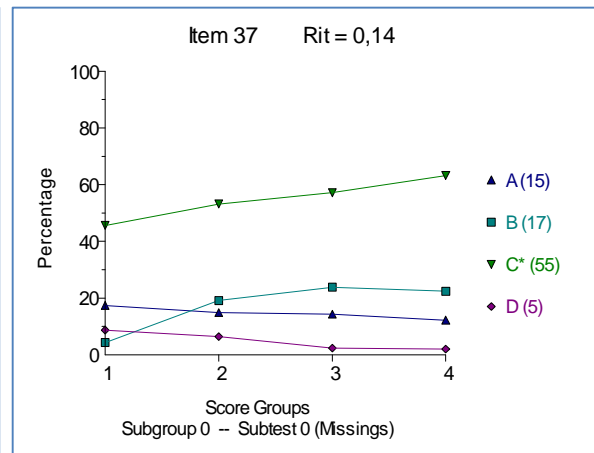
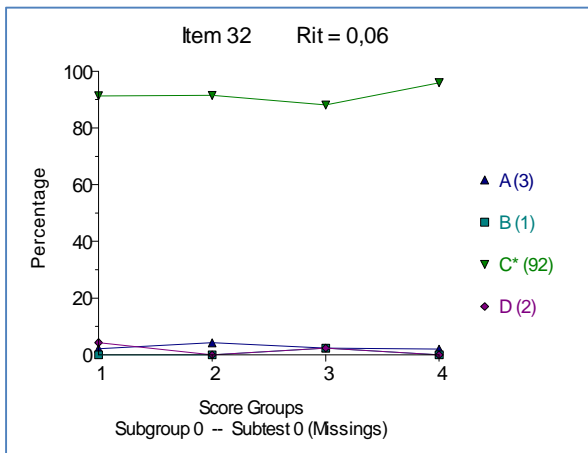
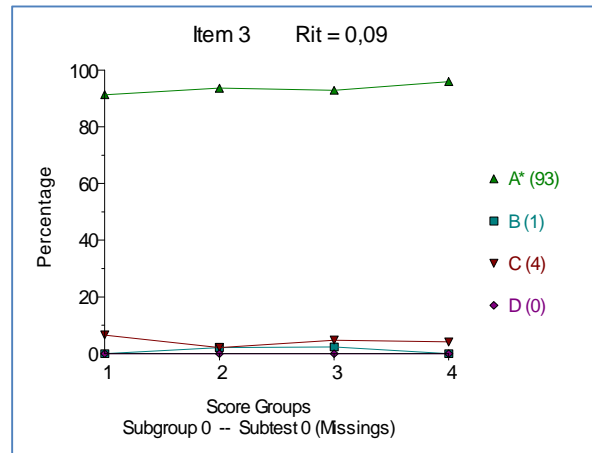
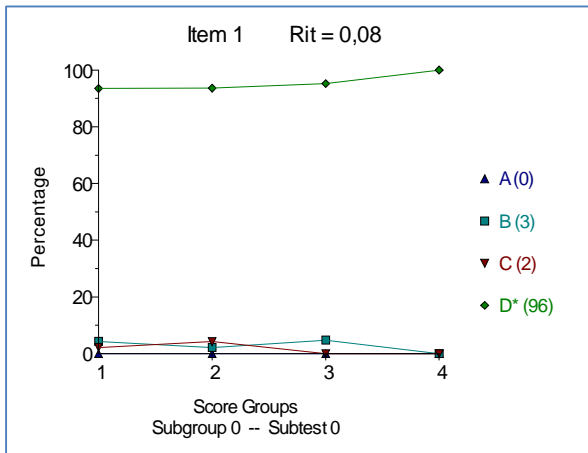


Table B.5 Poor or pathological items in PCD-Matemática

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
1	0,96	0,08	0,05	A	This is poor because there is no REAL alternative for the correct answer
3	0,93	0,09	0,06	A	This is poor because there is no REAL alternative for the correct answer
32	0,92	0,06	0,03	AB	This is poor because there is no REAL alternative for the correct answer
37	0,55	0,14	0,09	ABD	This is poor because the BEST students are messing with B. Actually the B seems to be quite good option for a correct one. (Check the Key!)
47	0,3	0,14	0,09	AB	This is poor because there seems to be TWO correct answers (D and A)
50	0,76	0,13	0,09	A	This is poor because the POOREST find the correct answer too easily
57	0,19	0,00	-0,04	ABCD	This is pathological because the BEST ones do not find the correct alternative. The correct alternative seems to be C (not B)?

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high



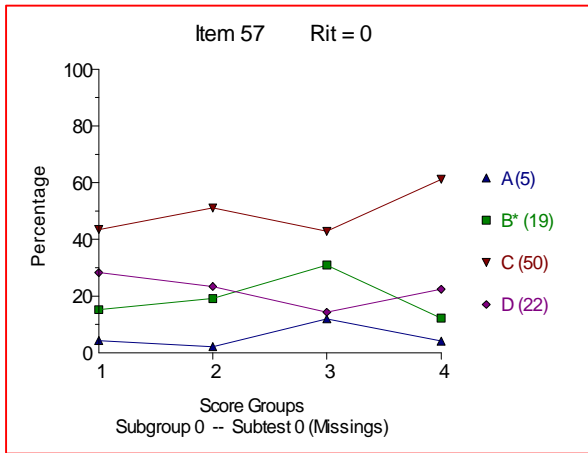
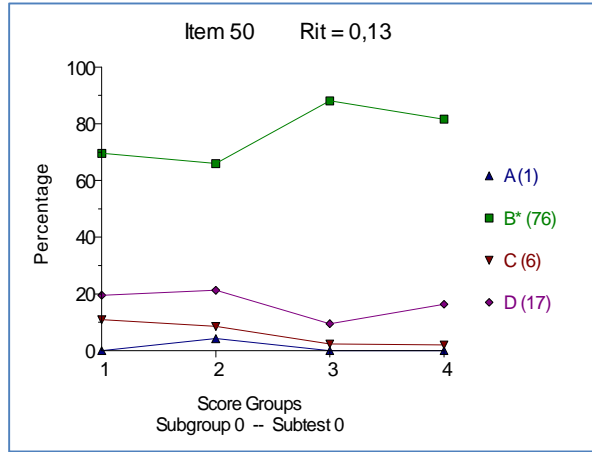
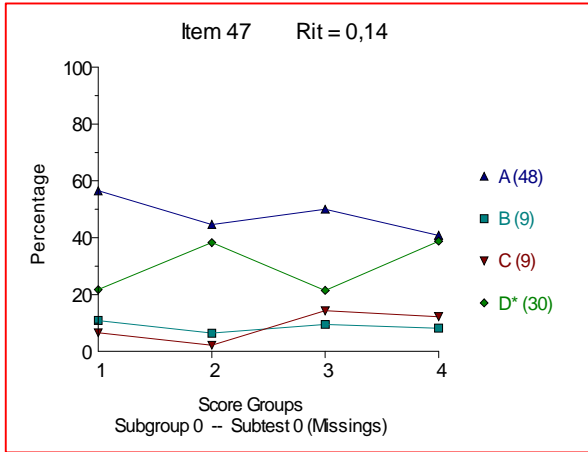
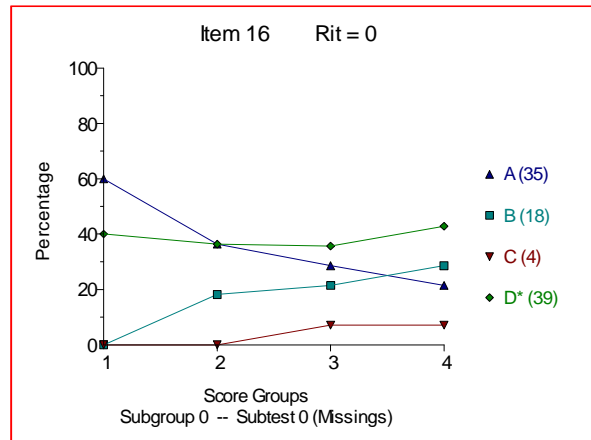
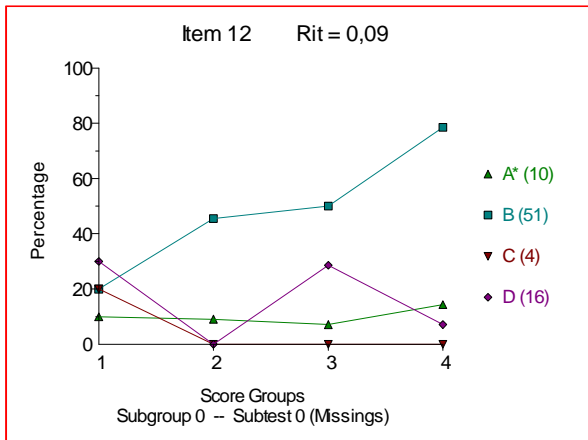
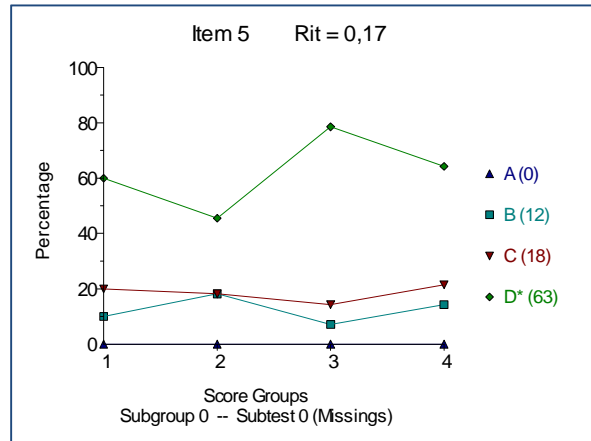
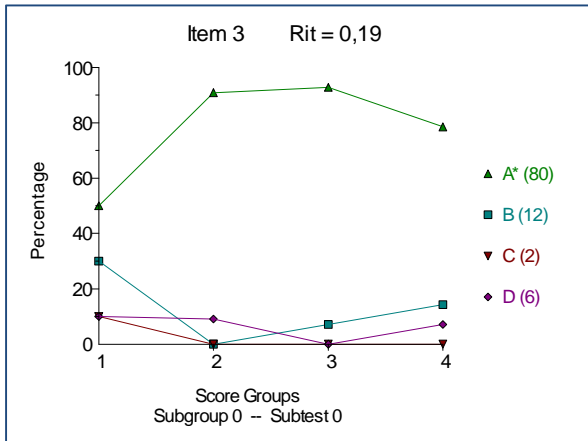
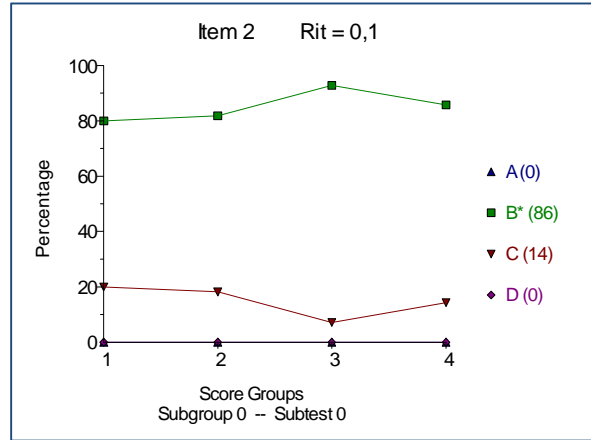
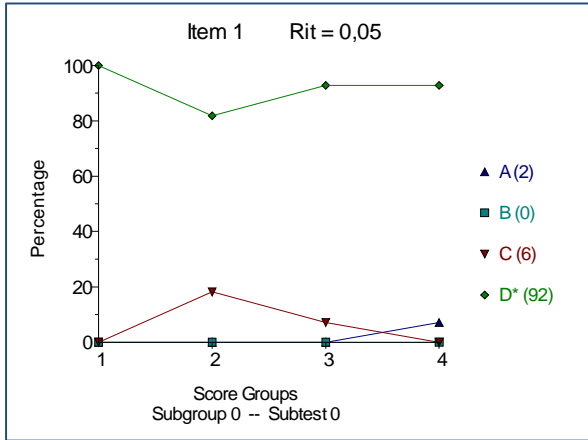
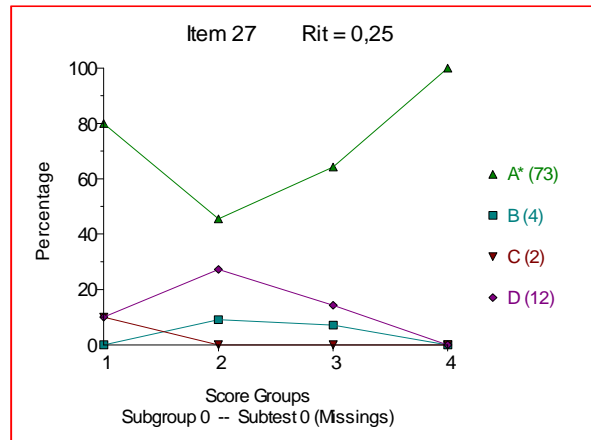
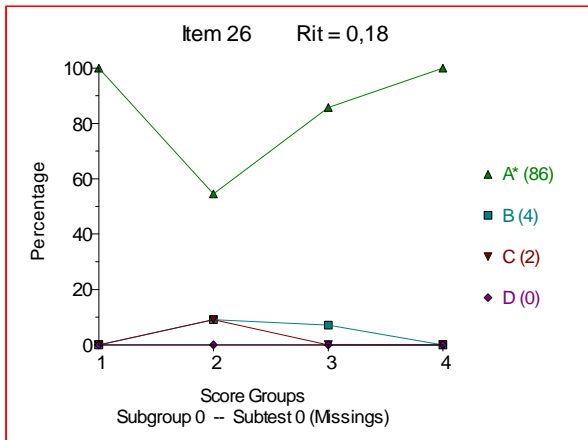
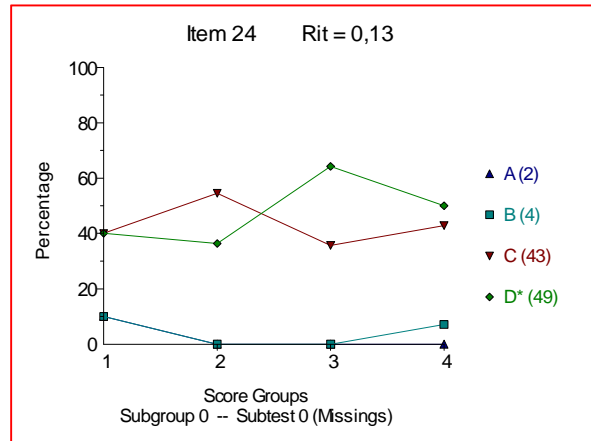
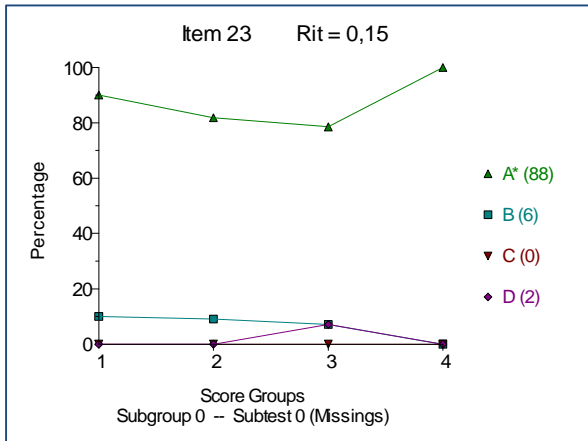
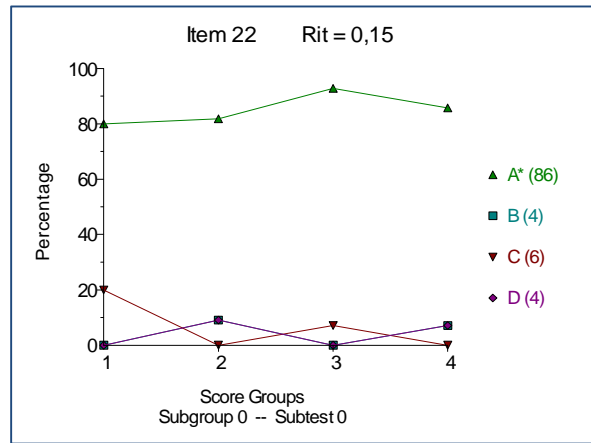
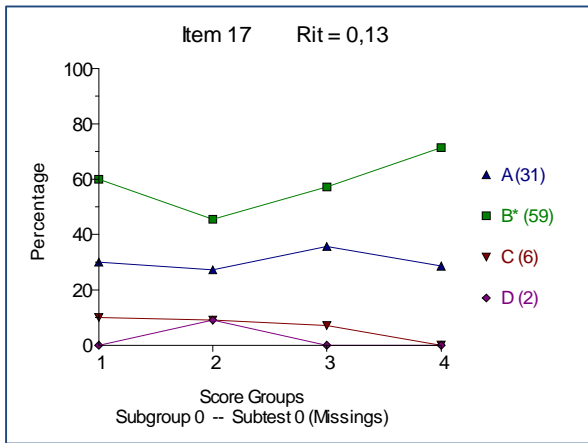


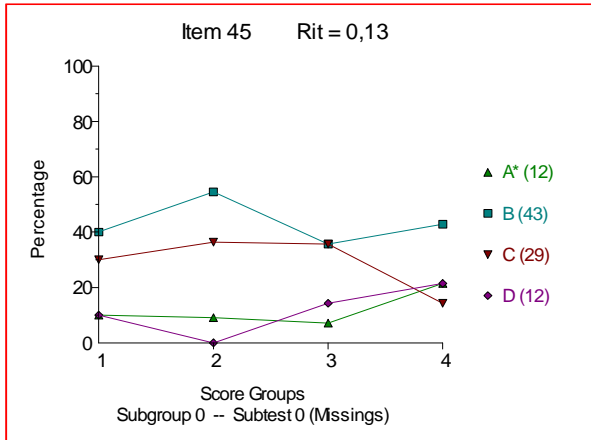
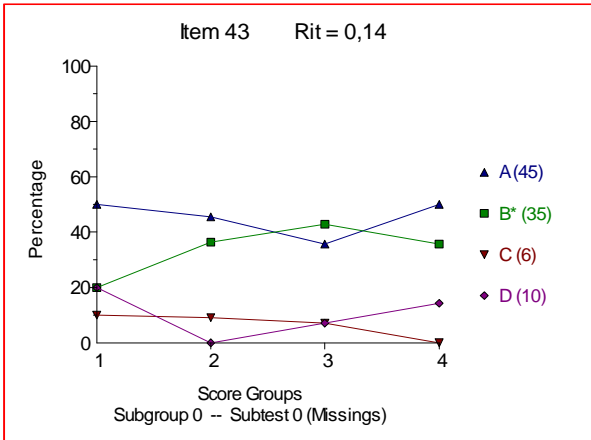
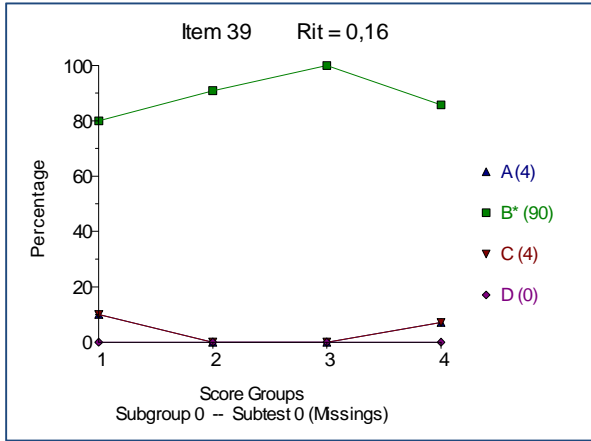
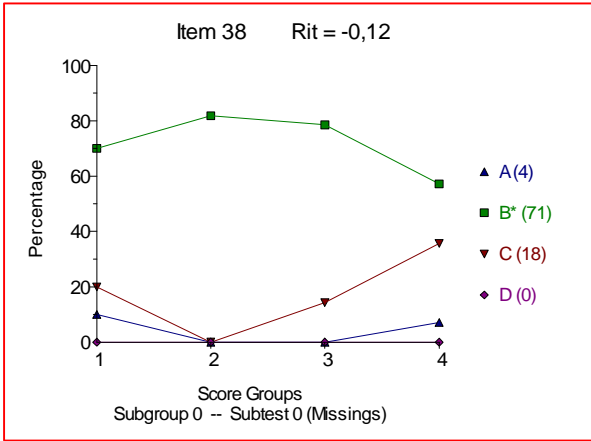
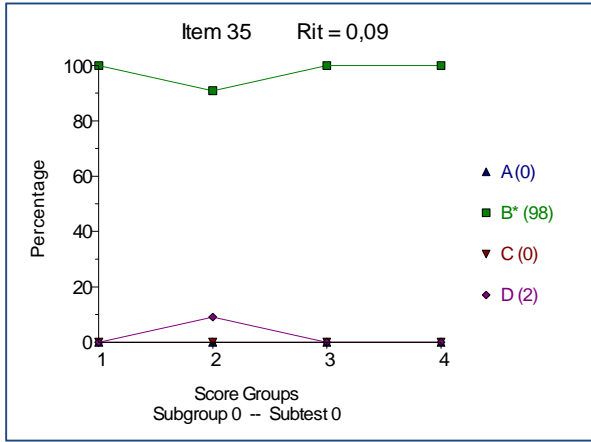
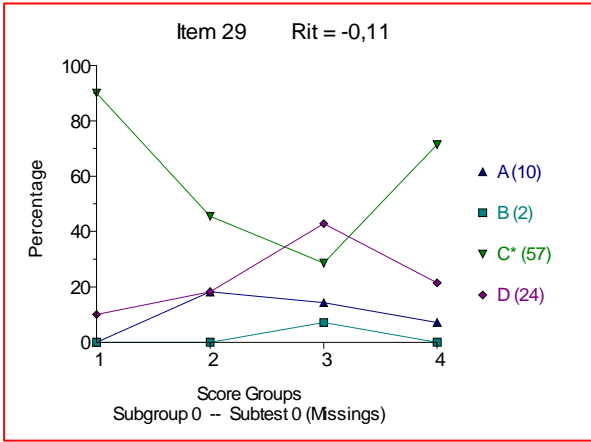
Table B.6 Poor or pathological items in PCP-Quimica

item nr.	p	Rit	Rir	Flag code ¹	Graphical analysis
1	0,92	0,05	0,01	ABD	There is no REAL alternative for the correct answer AND the WEAKEST find the correct answer too easily
2	0,86	0,1	0,05	A	There is no REAL alternative for the correct answer
3	0,8	0,19	0,13	A	The BEST students are messing with B. Sleepiness. Otherwise OK.
5	0,63	0,17	0,11	A	The BEST students are messing with C and B and the WEAKEST find the correct answer too easily . High Guessing.
12	0,1	0,09	0,05	ABD	This is pathological because there seems to be wrong key. Check the key. Definitely B is the correct one! Maybe B?
16	0,39	0	-0,07	ABCD	The WEAKEST find the correct answer too easily and the BEST are distracted to B. Check the key. B correct?
17	0,59	0,13	0,06	A	There seems to be TWO correct answers (B and A) and the WEAKEST find the correct answer too easily. High Guessing.
22	0,86	0,15	0,1	A	There is no REAL alternative for the correct answer
23	0,88	0,15	0,11	A	There is no REAL alternative for the correct answer AND the WEAKEST find the correct answer too easily
24	0,49	0,13	0,06	A	There seems to be TWO correct answers (C and D)
26	0,86	0,18	0,13		This is pathological because the WEAKEST find the correct answer too easily. High Guessing.
27	0,73	0,25	0,19		This is pathological because the WEAKEST find the correct answer too easily. High Guessing.
29	0,57	-0,11	-0,18	ABCD	This is pathological because the WEAKEST find the correct answer too easily. There seems to be another correct answer, D
35	0,98	0,09	0,07	A	There is no REAL alternative for the correct answer and the WEAKEST find the correct answer too easily . Guessing. Just too easy item.
38	0,71	-0,12	-0,18	ABCD	This is pathological because the WEAKEST find the correct answer too easily and the BEST ones are messing with C. Check the key! There seems to be another correct answer, C.
39	0,9	0,16	0,12	A	There is no REAL alternative for the correct answer
42	0,37	0,23	0,16	D	
43	0,35	0,14	0,08	A	There seems to be TWO correct answers (A and B)
45	0,12	0,13	0,09	ABD	There seems to be SEVERAL or NO correct answers and the BEST ones are messing with B
46	0,84	0,18	0,13	AD	There is no REAL alternative for the correct answer and the BEST ones are distracted to D
47	0,76	0,07	0,01	AB	There is no REAL alternative for the correct answer and the WEAKEST find the correct answer too easily.
48	0,71	0,19	0,13	A	no problem
50	0,47	0,5	0,45	D	
51	0,45	0,17	0,1	A	There seems to be TWO correct answers (B and A) and the BEST ones are messing with A
52	0,2	0,35	0,3	D	
54	0,84	0,17	0,12	A	There is no REAL alternative for the correct answer and the WEAKEST find the correct answer too easily . Guessing.
57	0,45	0,12	0,05	AB	There seems to be SEVERAL alternatives for the correct answer and the BEST ones are distracted to D and the WEAKEST find the correct answer too easily. High Guessing.
59	0,53	0,26	0,19	BD	

1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high







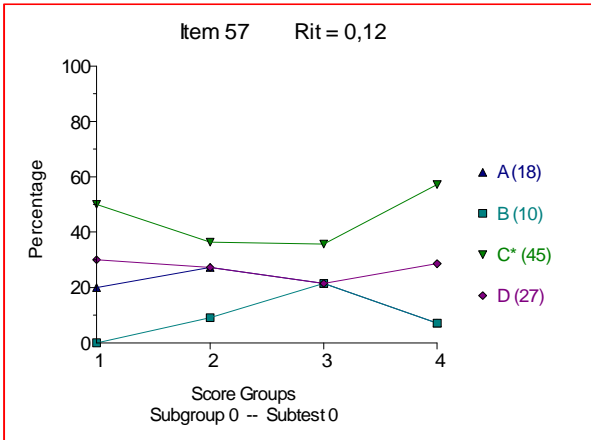
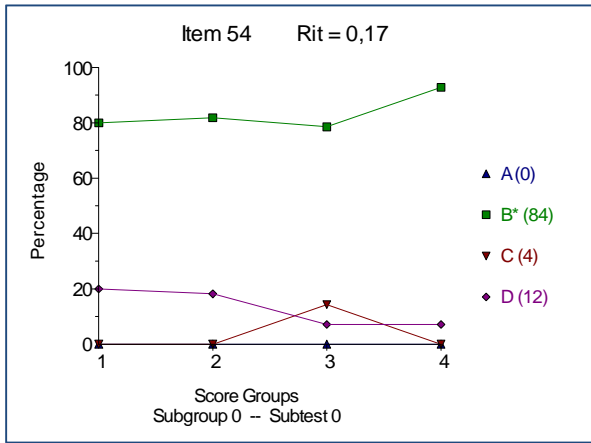
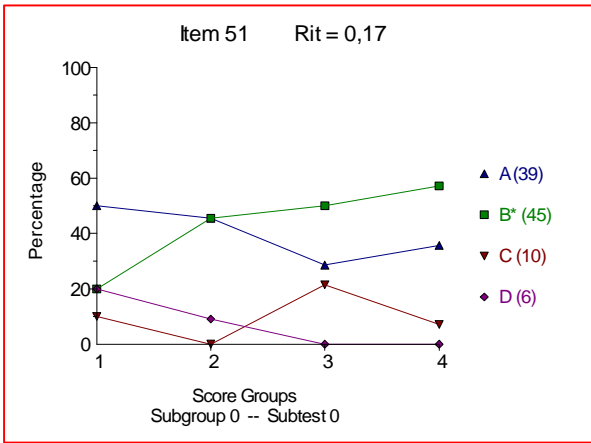
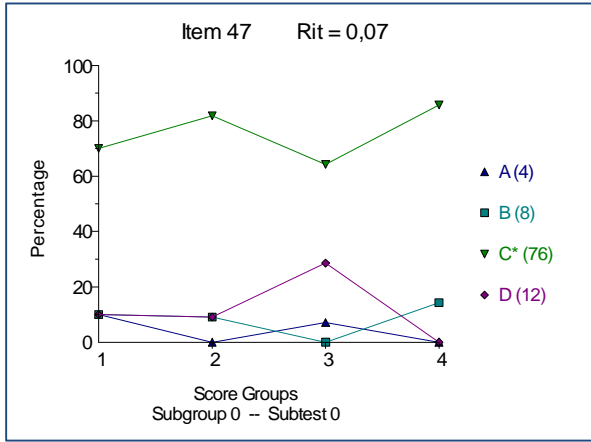
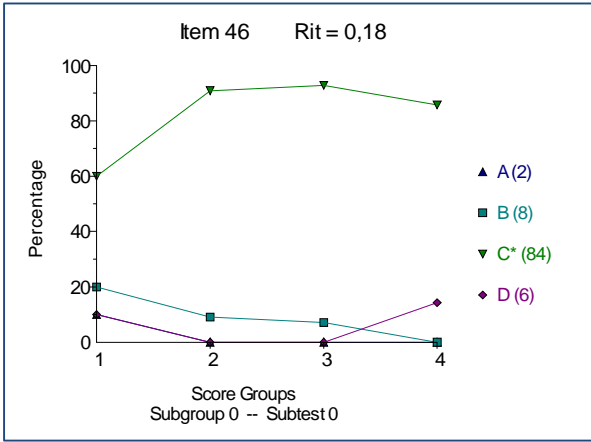
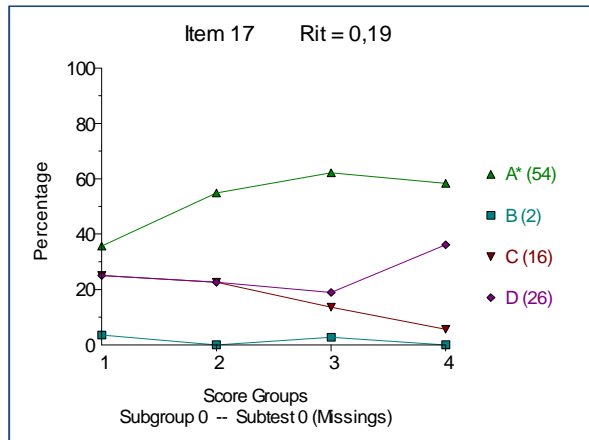
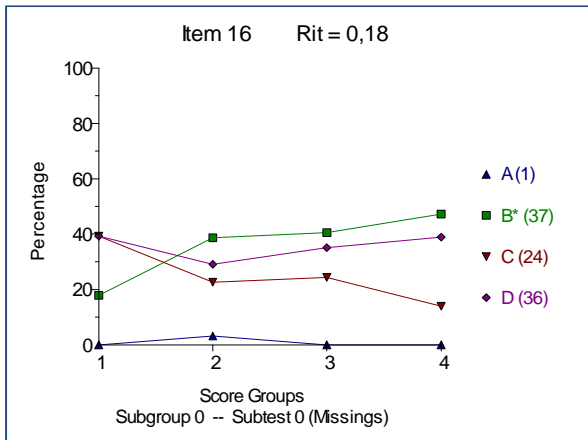
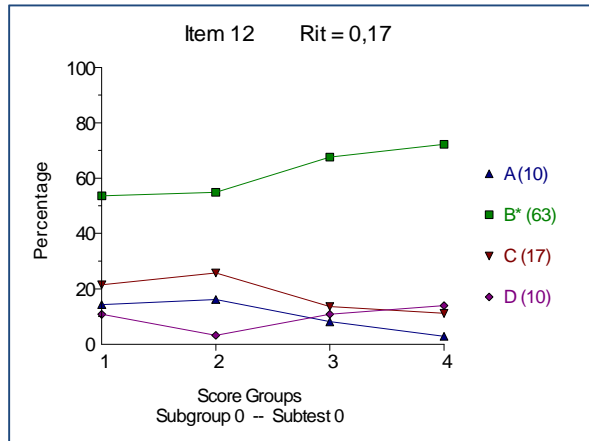
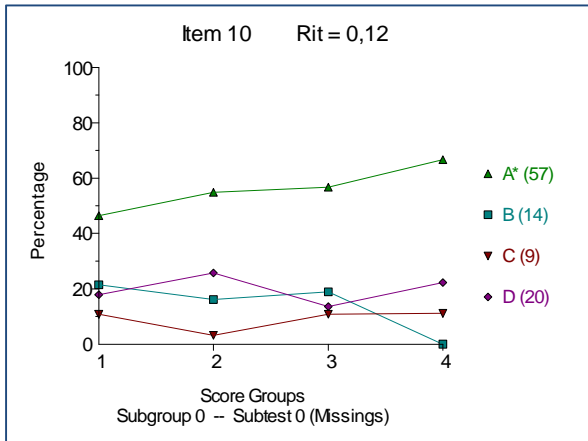
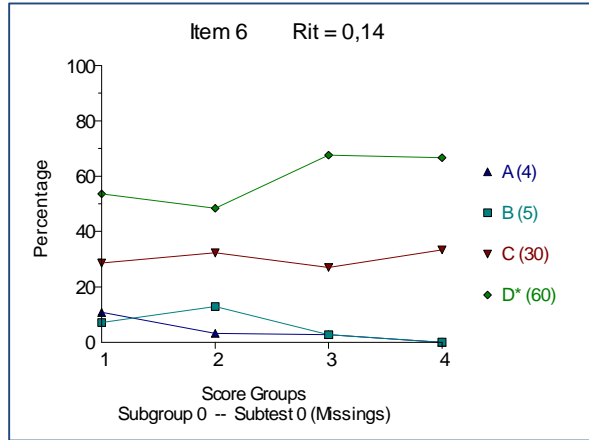
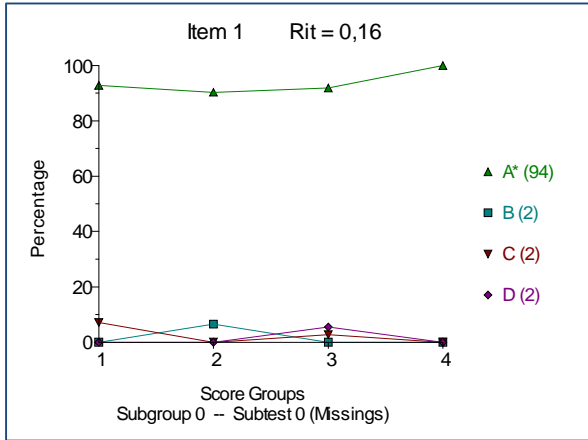
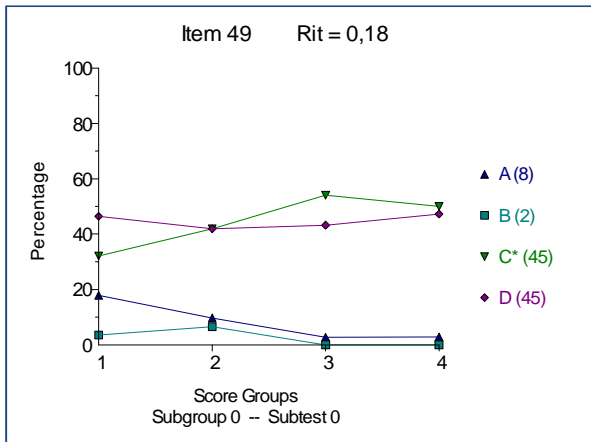
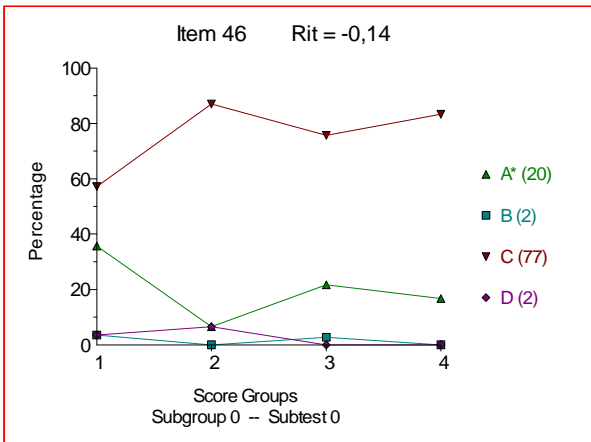
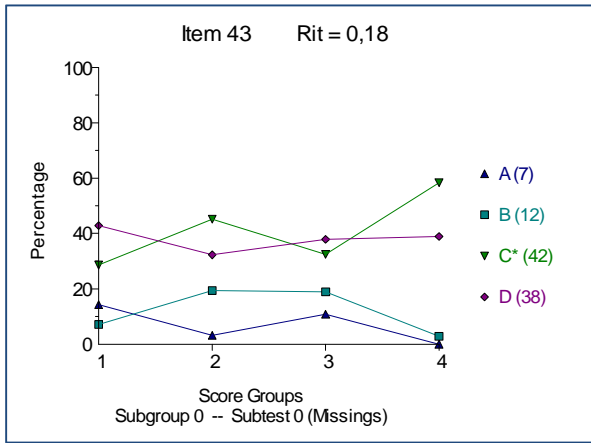
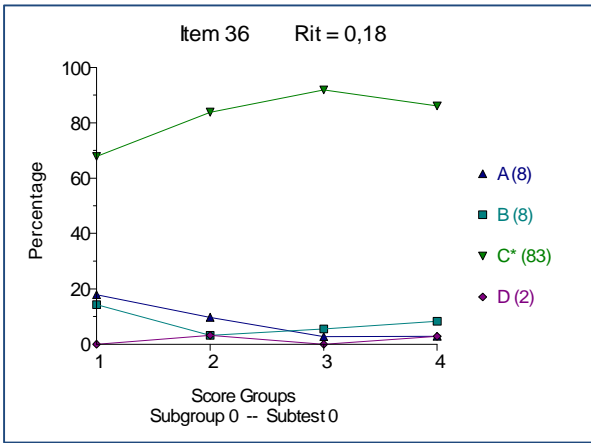
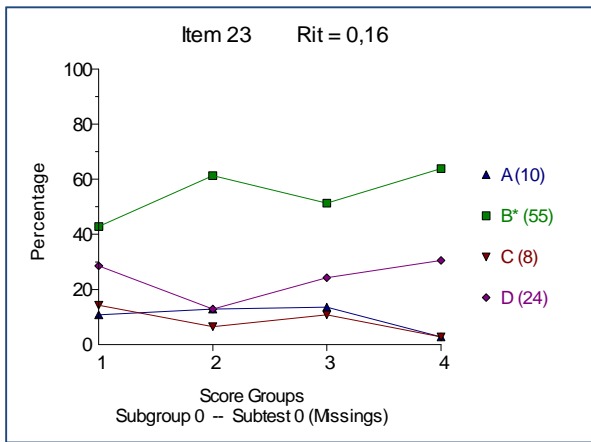
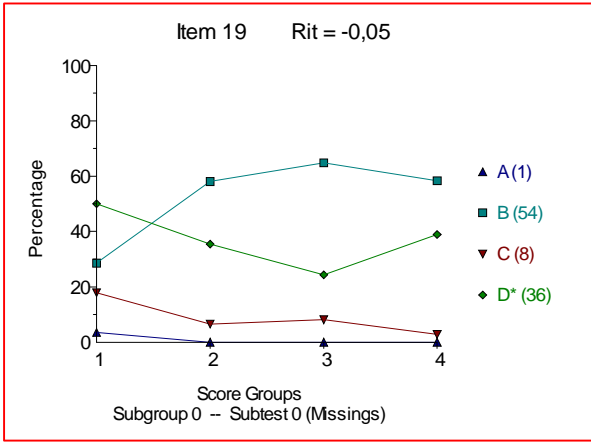


Table B.7A Poor or pathological items in PCD-Historia Version A

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
1	0,94	0,16	0,12	ABC	There is no REAL alternative for the correct answer
5	0,83	0,22	0,17	ABD	No problem
6	0,6	0,14	0,06	A	There seems to be TWO alternatives for the correct answer
7	0,48	0,46	0,4	D	
10	0,57	0,12	0,05	A	There is no REAL alternative for the correct answer
12	0,63	0,17	0,1	A	There is no REAL alternative for the correct answer, and the weakest students find the correct alternative too easily
16	0,37	0,18	0,11	A	There seems to be TWO alternatives for the correct answer (D and B)
17	0,54	0,19	0,12		The BEST students seems to distracted to D
19	0,36	-0,05	-0,12	A	This is pathological because the POOREST guess the correct, alternative and the BEST ones are distracted to B (instead of D)
23	0,55	0,16	0,08	A	There seems to be TWO alternatives for the correct answer (B and D)
36	0,83	0,18	0,12		There is no REAL alternative for the correct answer
43	0,42	0,18	0,11	ABCD	There seems to be TWO alternatives for the correct answer (C and D)
46	0,2	-0,14	-0,19	A	This is pathological because the BEST students do not know the correct answer. Check the key - C could be correct.
49	0,45	0,18	0,11	A	The BEST students do not find the correct alternative and because the weakest students find the correct alternative too easily
50	0,66	0,10	0,03	A	There is no REAL alternative for the correct answer and because the BEST students do not find the correct alternative
51	0,44	0,10	0,03	A	The BEST students do not find the correct alternative and because the weakest students find the correct alternative too easily
56	0,81	0,21	0,16	BD	
57	0,51	0,17	0,09	A	There seems to be several good options for the best students
59	0,64	0,18	0,11	A	The BEST students do not find the correct alternative and because the weakest students find the correct alternative too easily
60	0,04	0,02	-0,01	A	There is no REAL alternative for the correct answer. Check the Key. May be C?

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high





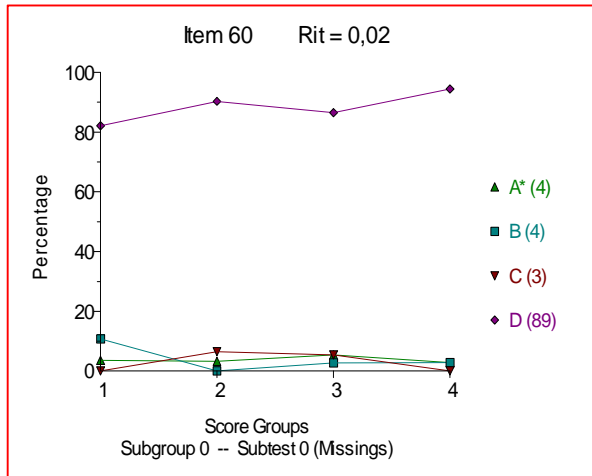
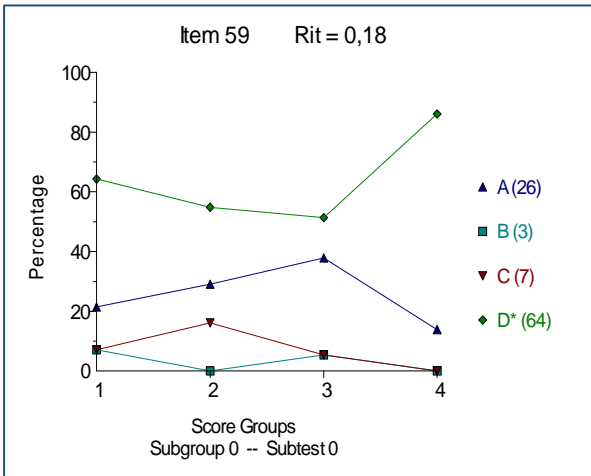
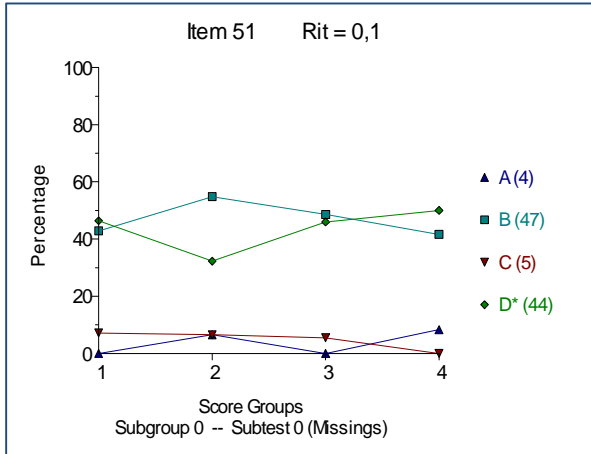
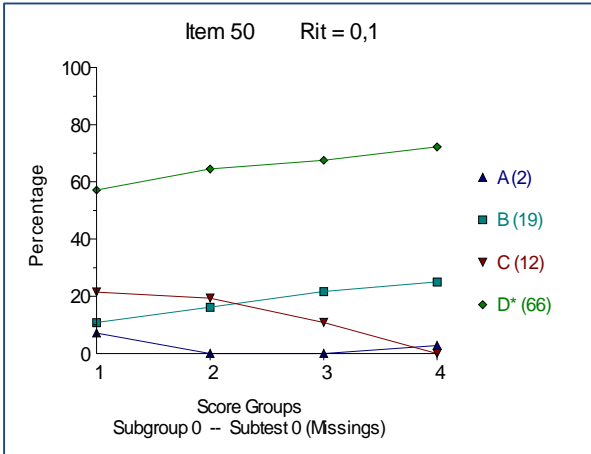
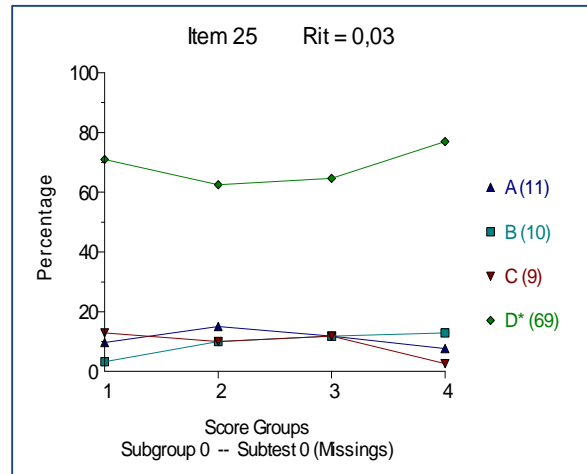
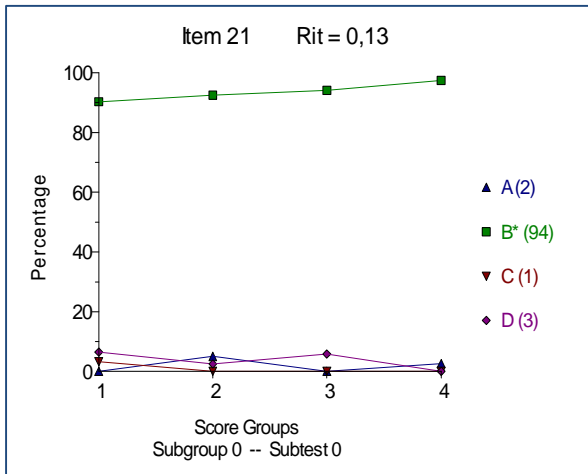
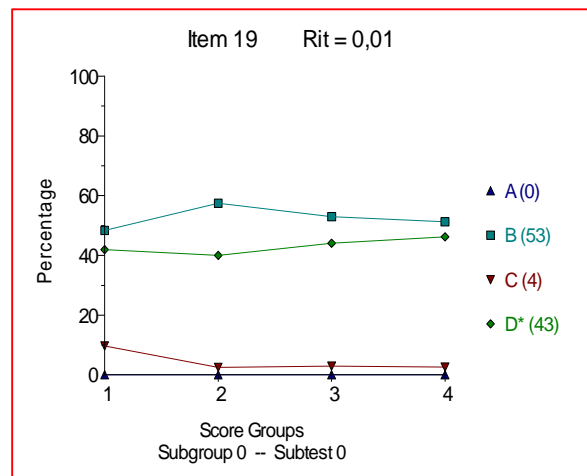
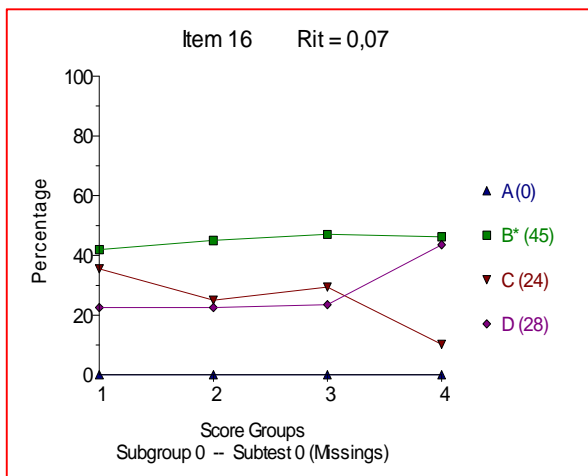
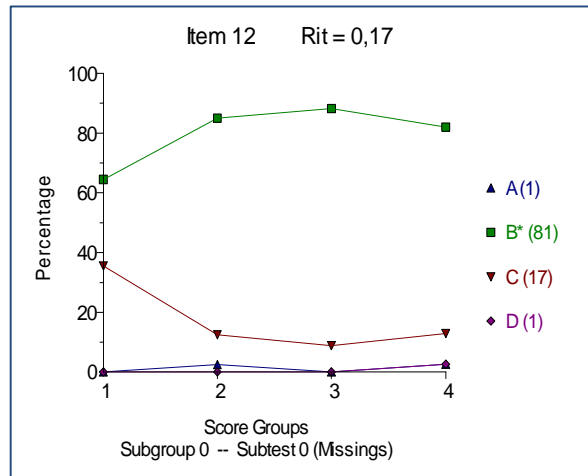
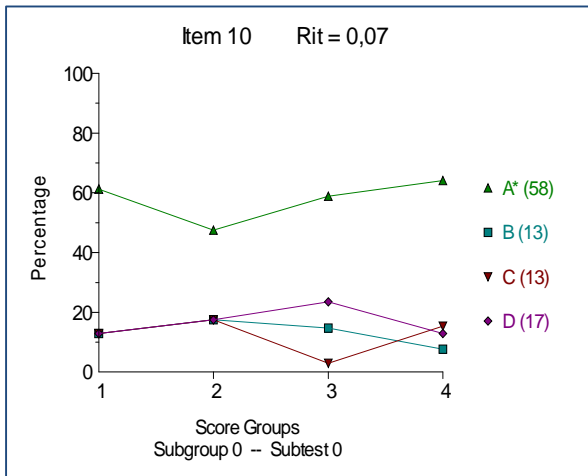
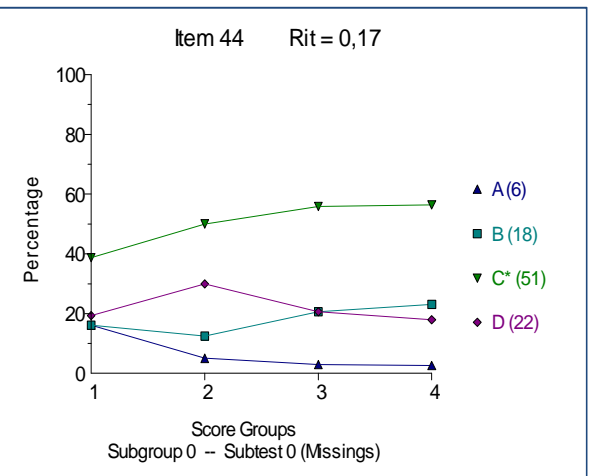
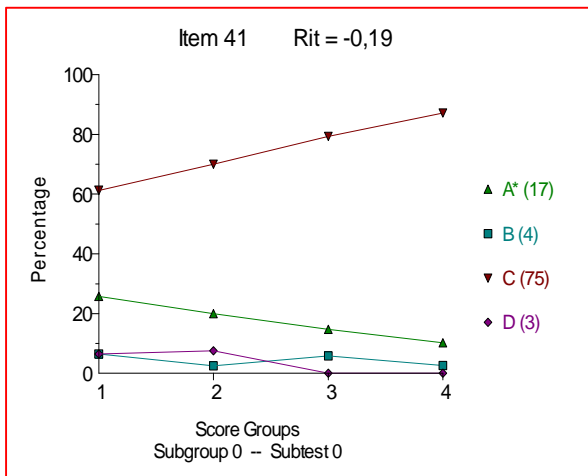
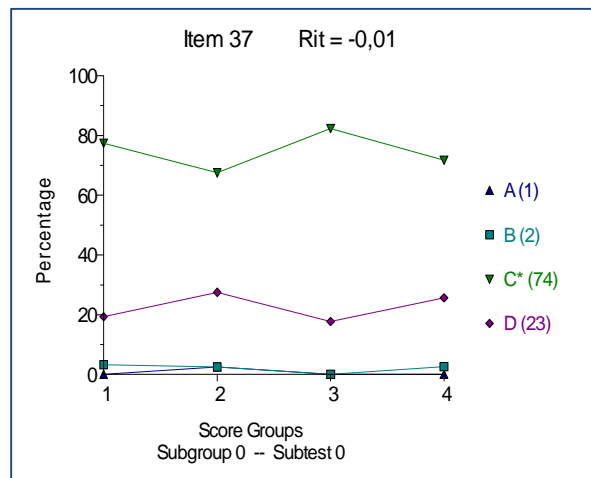
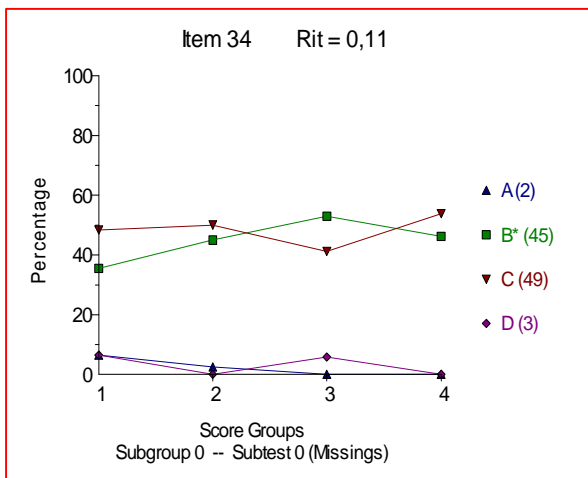
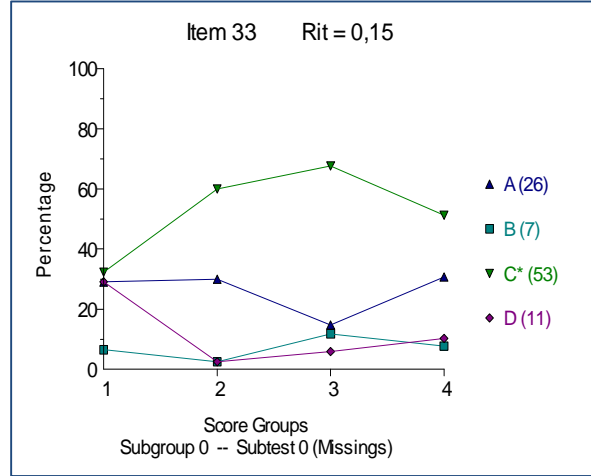
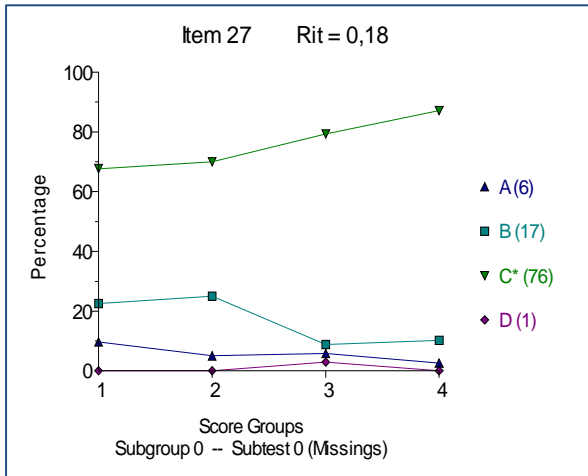


Table B.7B Poor or pathological items in PCD-Historia Version B

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
7	0,85	0,19	0,13	A	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
10	0,58	0,07	0,00	ABC	There is no REAL alternative for the correct answer and because of high guessing parameter
12	0,81	0,17	0,11	AD	There is no REAL alternative for the correct answer
16	0,45	0,07	-0,01	ABCD	There seems to be TWO correct answers (B and D). The weakest students find the correct alternative too easily and the BEST students are messing with D
19	0,43	0,01	-0,06	ABCD	There seems to be TWO correct answers (B and D)
21	0,94	0,13	0,10		There is no REAL alternative for the correct answer
25	0,69	0,03	-0,04	ABCD	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
27	0,76	0,18	0,11	A	There is no REAL alternative for the correct answer
33	0,53	0,15	0,08	A	The BEST students are distracted by A
34	0,45	0,11	0,04	AB	There are TWO correct alternatives (C and B)
37	0,74	-0,01	-0,07	ABC	The weakest students find the correct alternative too easily, and he BEST ones are messing with D. TWO correct?
41	0,17	-0,19	-0,24	ABCD	This is pathological because the BEST students do not find the correct alternative. Definitely A is not the correct answer. I'd guess C instead.
44	0,51	0,17	0,10	ABD	The BEST students are messing with B and D. No correct answer?
51	0,51	0,18	0,10	A	There are TWO correct alternatives (A and B)
59	0,92	0,16	0,12	A	There is no REAL alternative for the correct alternative

1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high





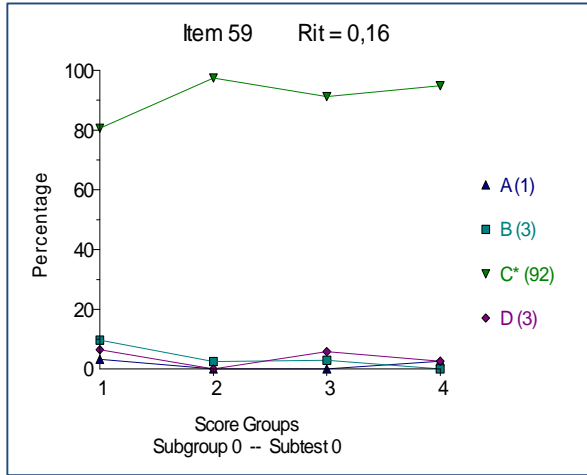
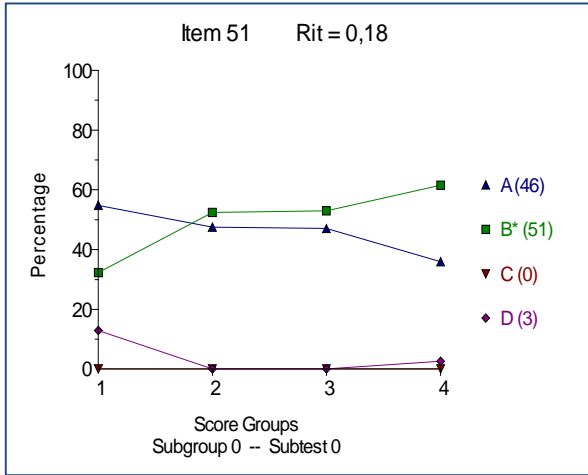
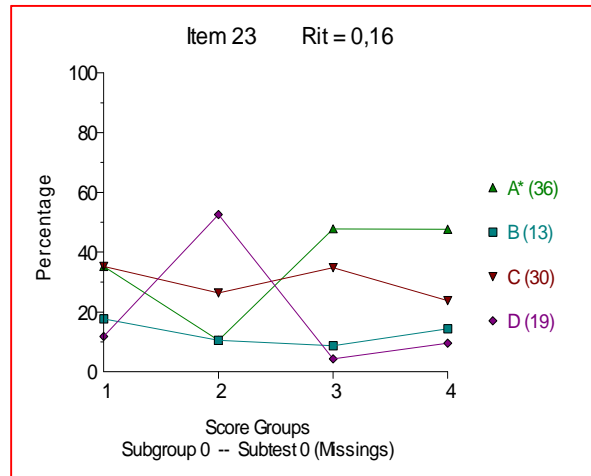
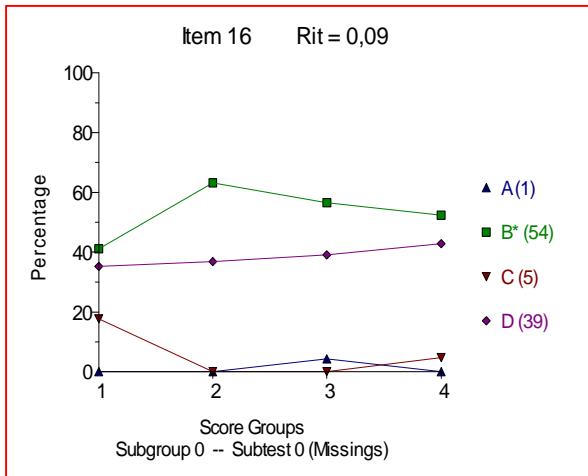
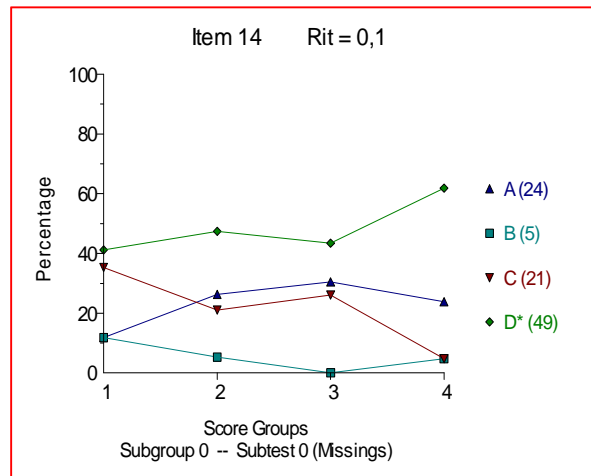
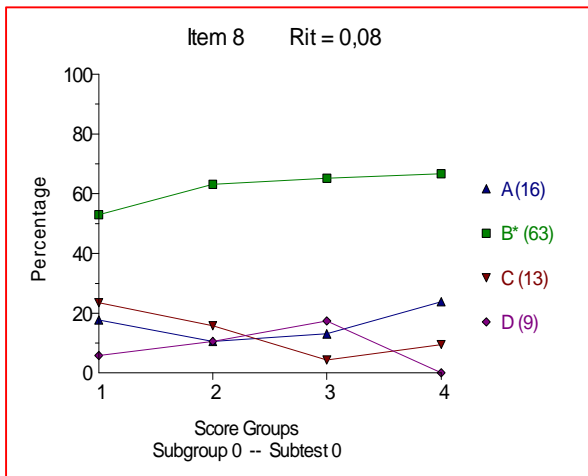
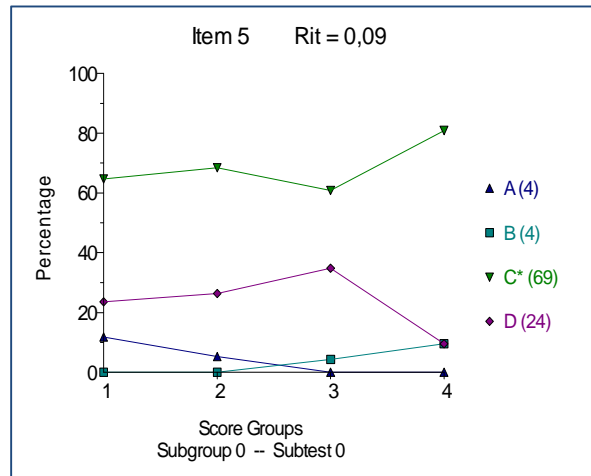
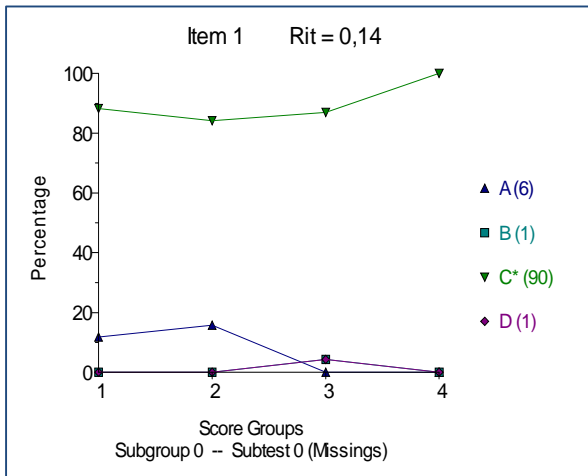
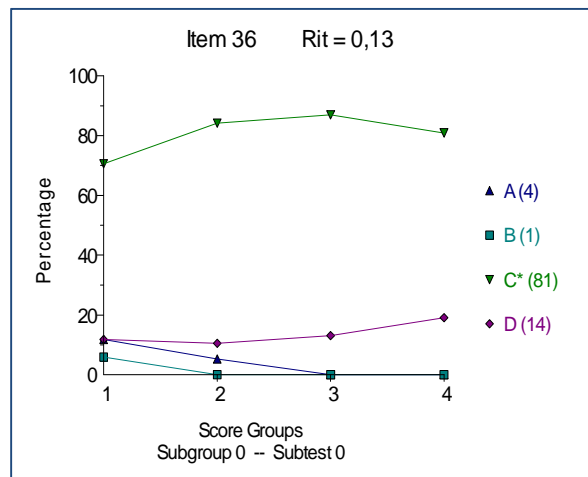
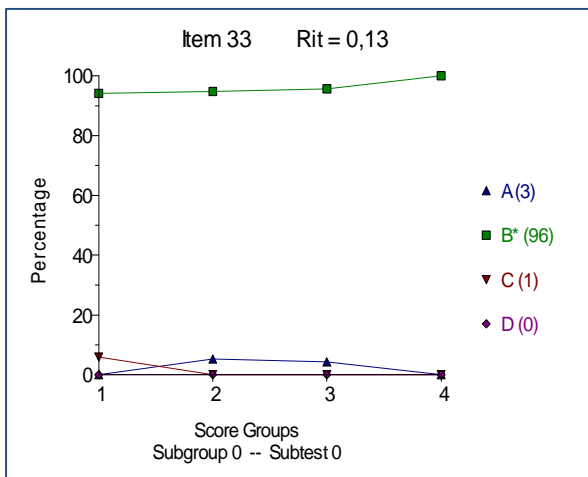
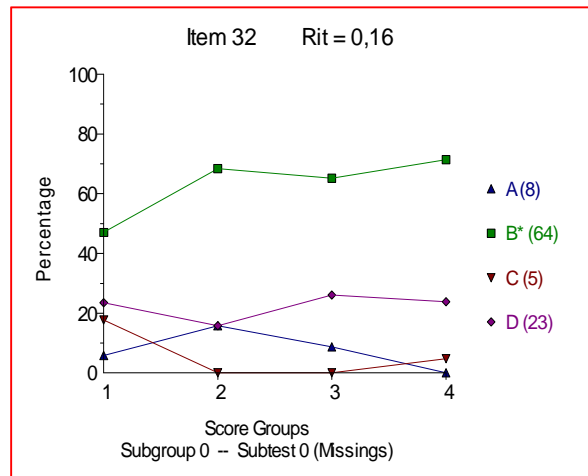
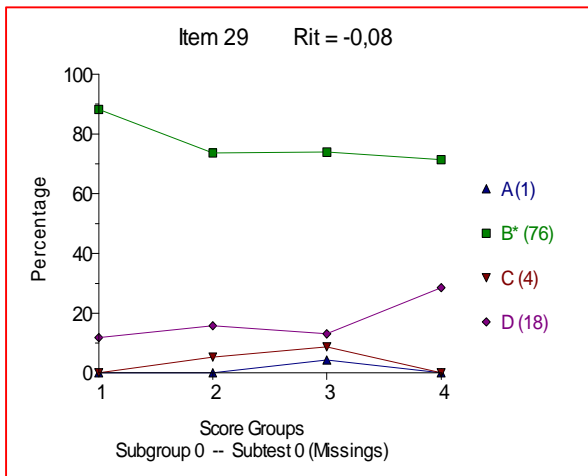
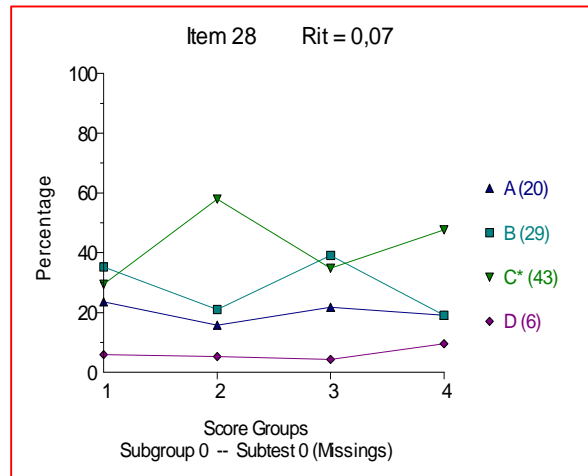
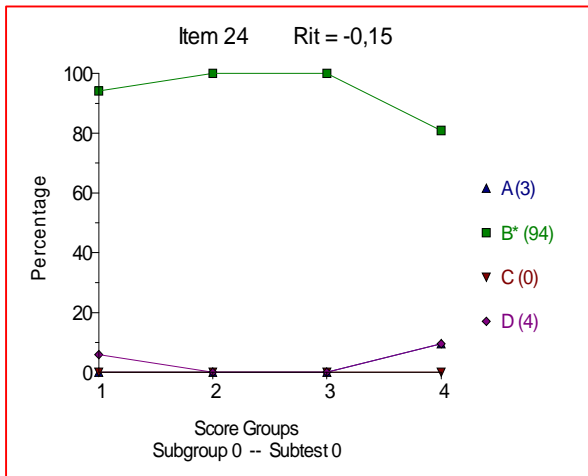


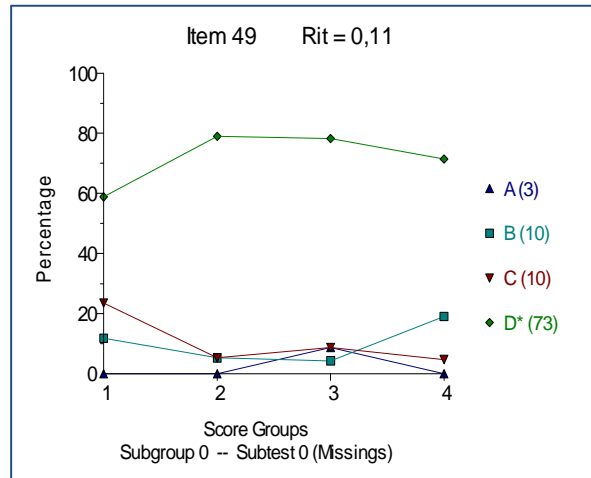
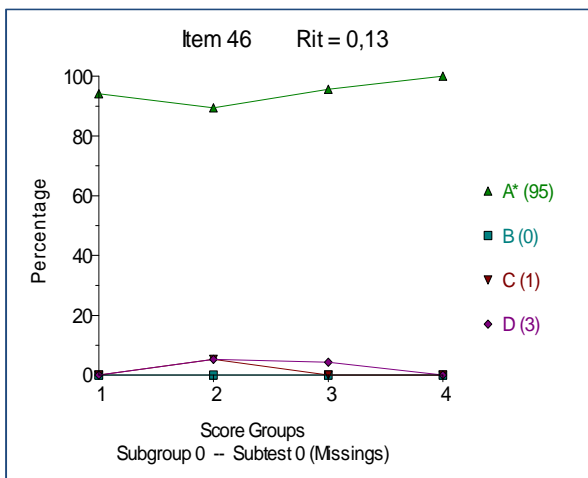
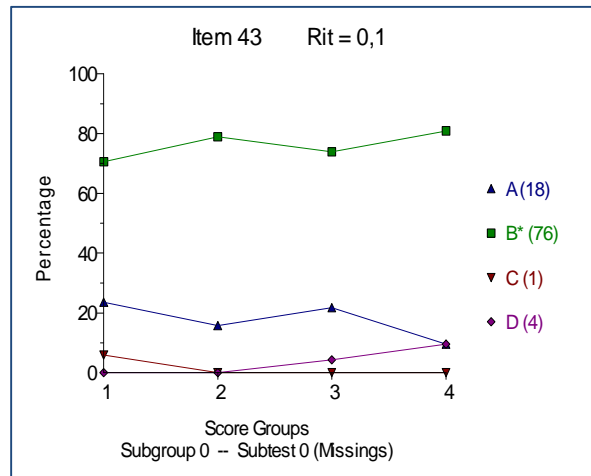
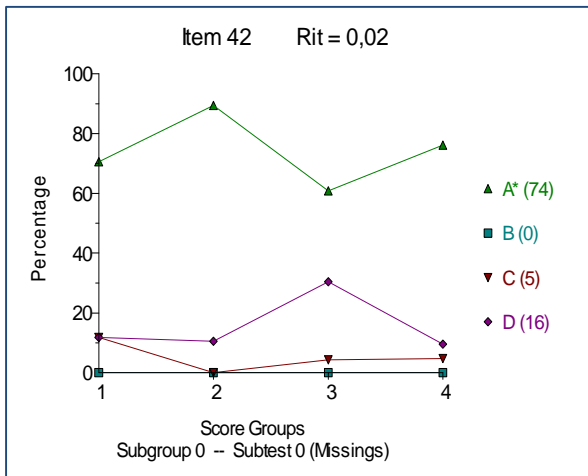
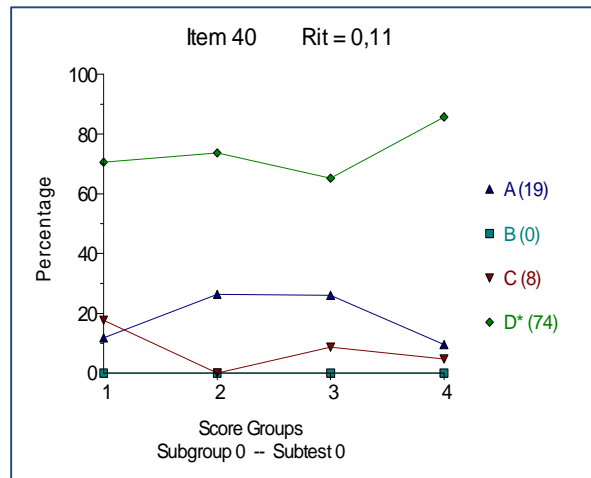
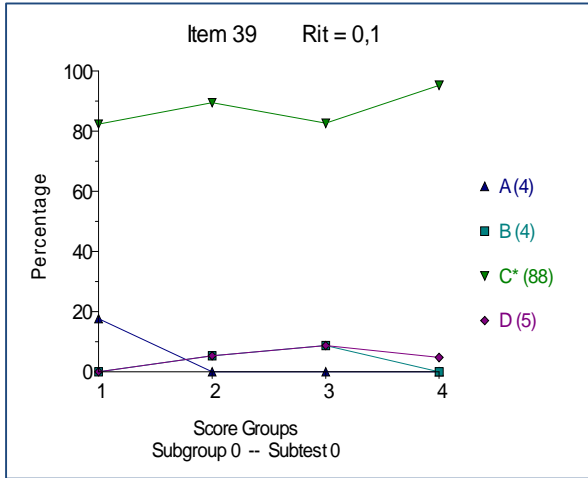
Table B.8A Poor or pathological items in PCD-Lenguaje Version A

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
1	0,9	0,14	0,09	A	There is no REAL alternative for the correct answer
5	0,69	0,09	0,01	ABD	This is poor because WEAKEST students find the correct answer too easily
8	0,63	0,08	0	ABD	This is poor because, for the BEST students, there seems to be TWO alternatives for the correct answer (B and A)
14	0,49	0,10	0,02	ABD	There is TWO alternatives for the correct answer for the BEST students (A and D), and the weakest students find the correct alternative too easily
16	0,54	0,09	0,01	AB	There seems to be TWO alternatives for the correct answer (D and B)
23	0,36	0,16	0,08	A	There seems to be NO correct answer
24	0,94	-0,15	-0,19	ABCD	This is pathological because the POOREST know the correct alternative and the BEST ones are distracted to D (instead of B)
26	0,56	0,36	0,29	D	
28	0,42	0,07	-0,01	ABC	There seems to be NO correct answer
29	0,76	-0,08	-0,15	ABCD	This is pathological because there seems to be TWO alternatives for the correct answer (D and B). Check the key! and because the POOREST know the correct
32	0,64	0,16	0,09	AB	There seems to be TWO alternatives for the correct answer (B and D)
33	0,96	0,13	0,1	A	There is no REAL alternative for the correct answer
36	0,81	0,13	0,06	ABD	The BEST students seem to be distracted to D. For them, there are TWO correct answers.
39	0,88	0,10	0,05	ABD	There is no REAL alternative for the correct answer
40	0,74	0,11	0,04	A	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
42	0,74	0,02	-0,05	ABC	The weakest students find the correct alternative too easily
43	0,76	0,10	0,03	ABD	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
46	0,95	0,13	0,09	A	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
49	0,72	0,11	0,04	ABD	This is poor because, for the BEST students, there seems to be TWO alternatives for the correct answer (B and D)
52	0,55	0,11	0,03	ABD	The BEST students seem to be distracted to D. For them, there are TWO correct answers (B and D)
53	0,99	0,16	0,14	A	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
57	0,57	0,14	0,07	ABD	The BEST students seem to be distracted to D. For them, there are TWO correct answers (A and D). Check the key!
59	0,55	0,14	0,06	ABD	The BEST students seem to be distracted to A. For them, there are TWO correct answers (A and C).
60	0,39	0,26	0,19	D	

1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high







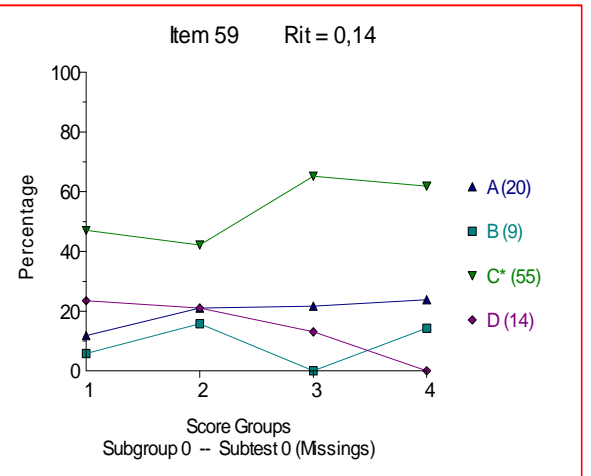
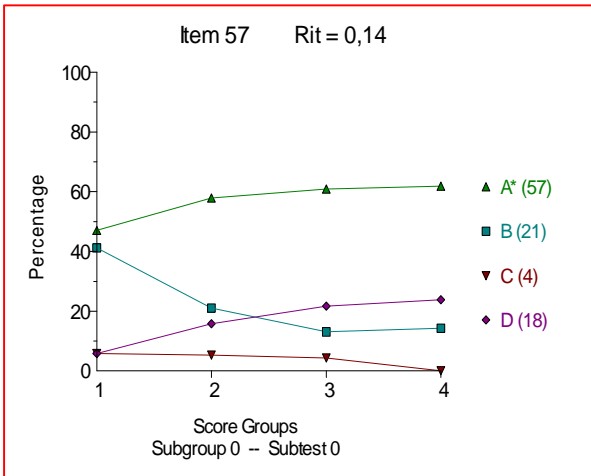
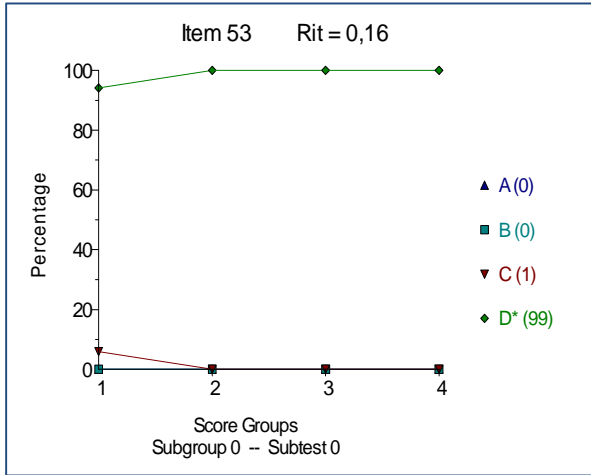
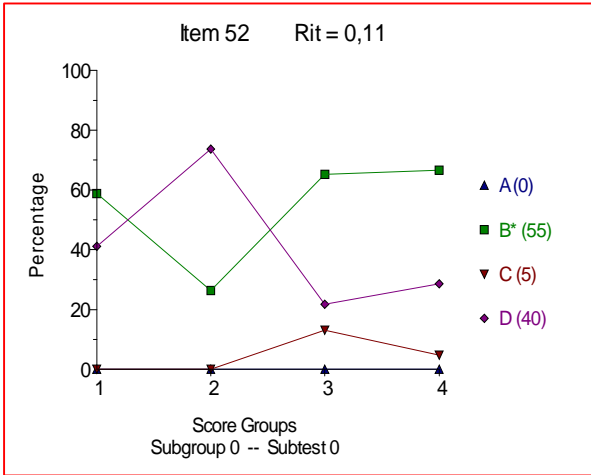
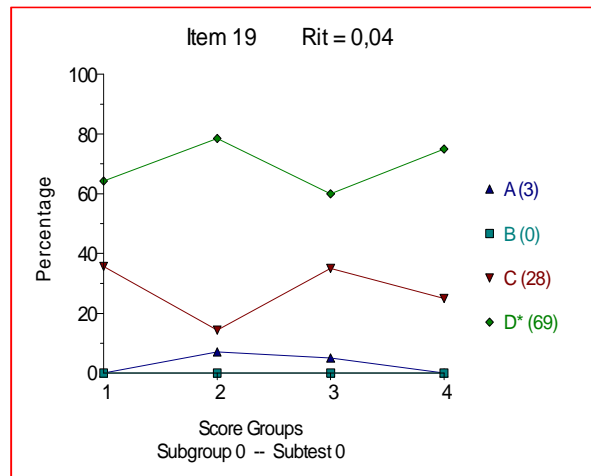
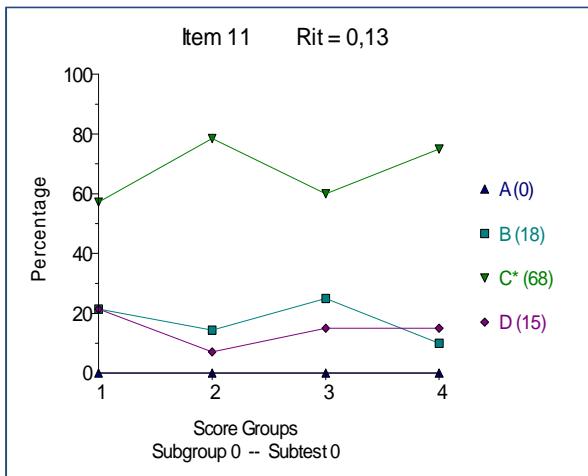
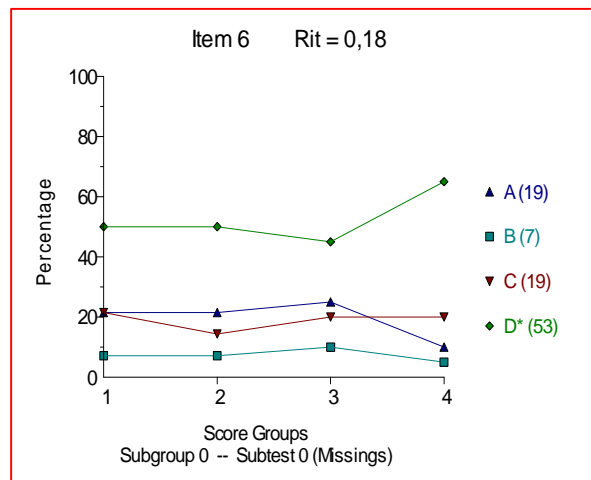
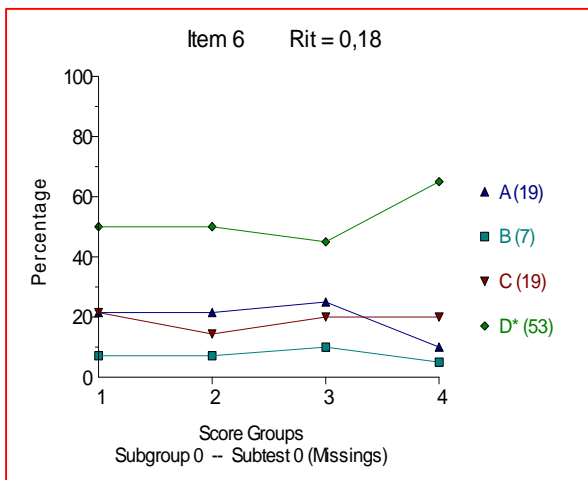
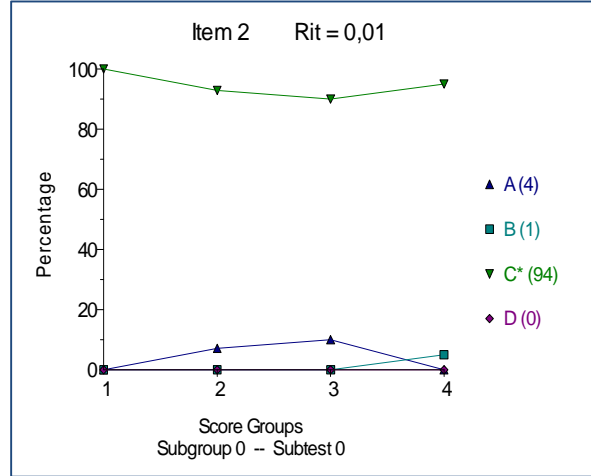
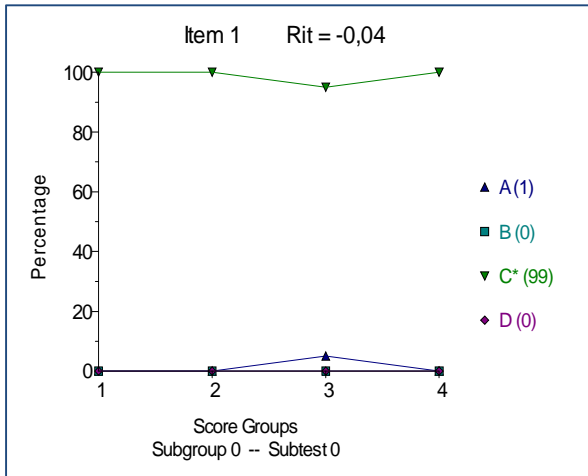
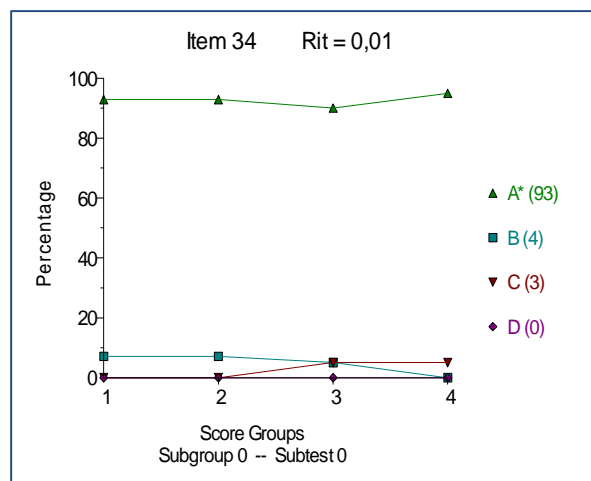
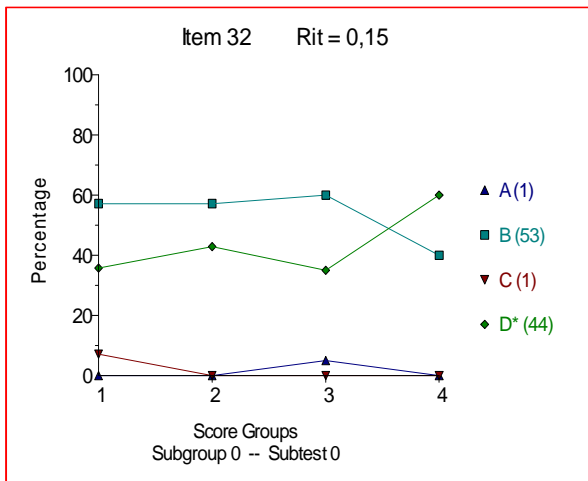
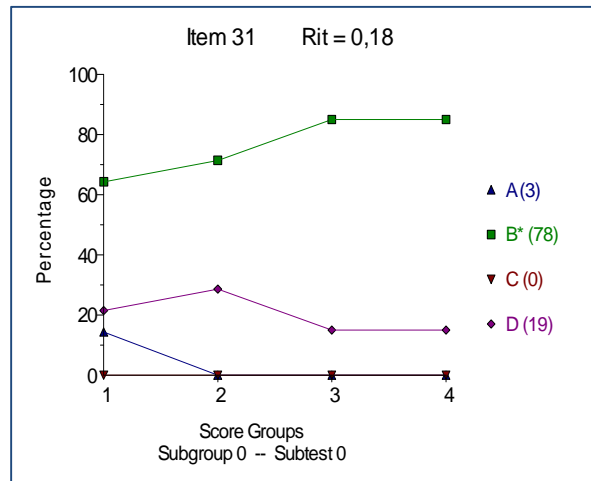
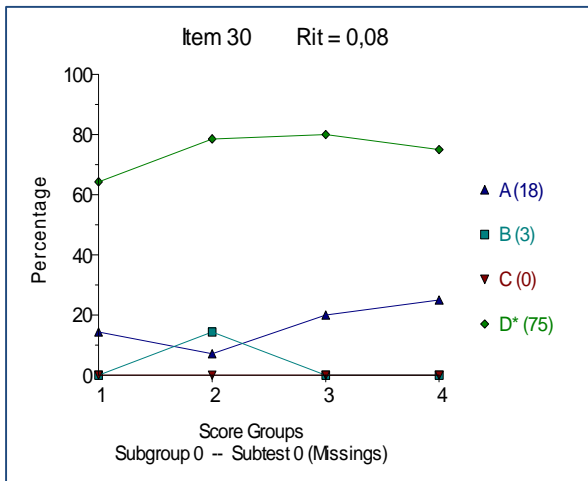
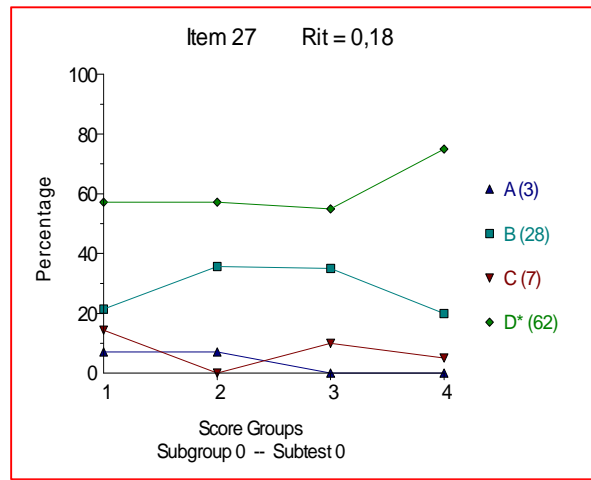
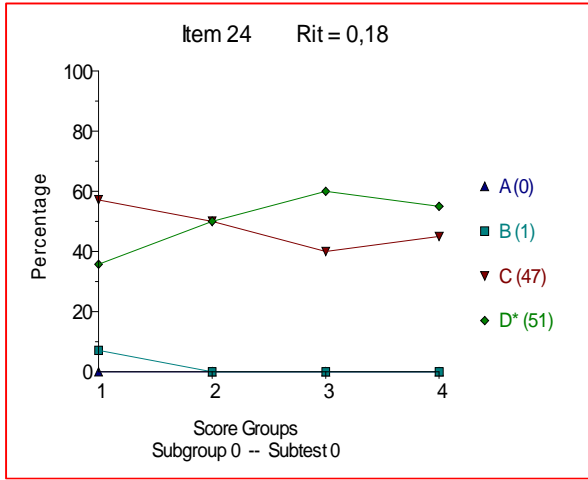


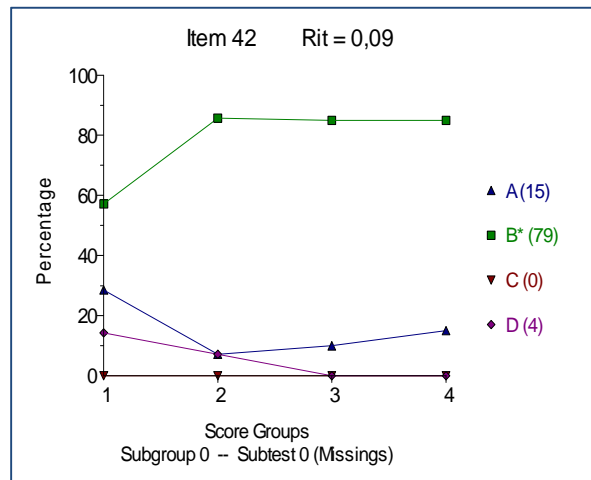
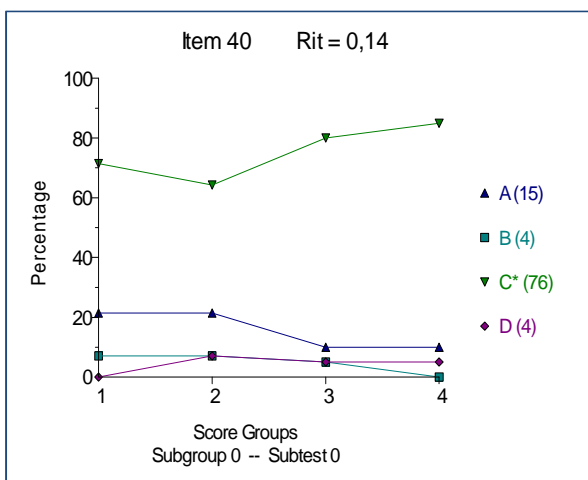
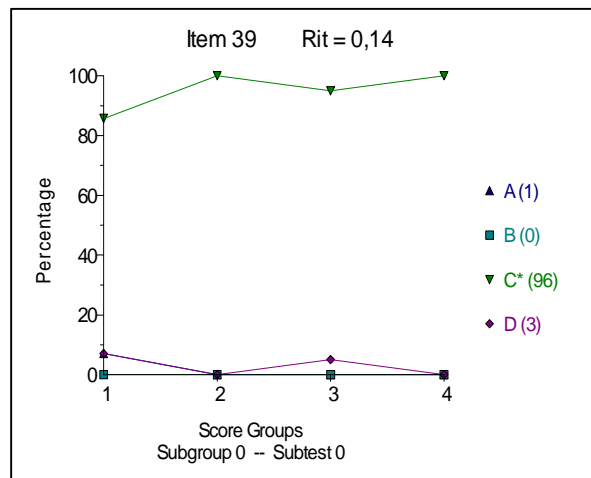
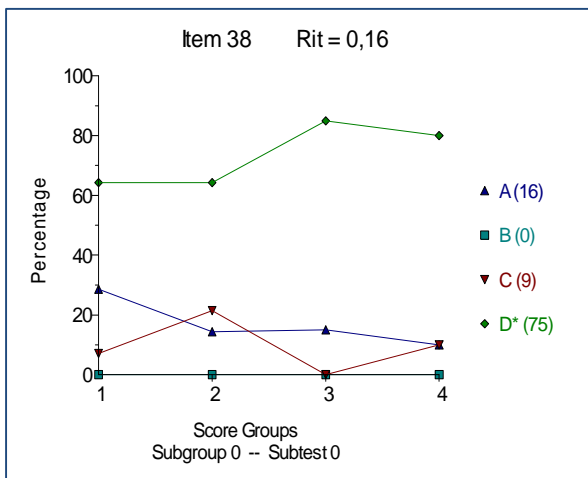
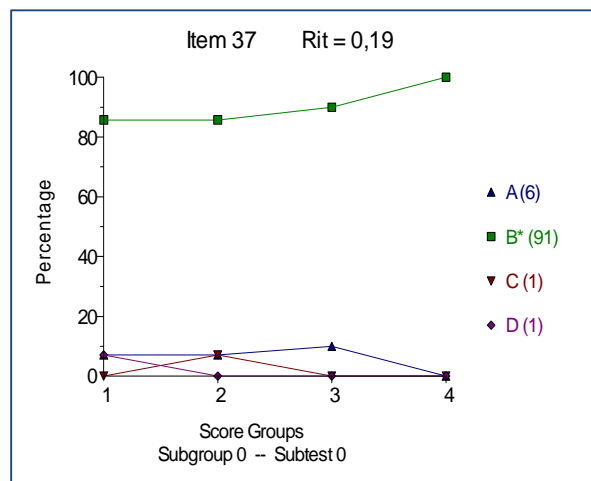
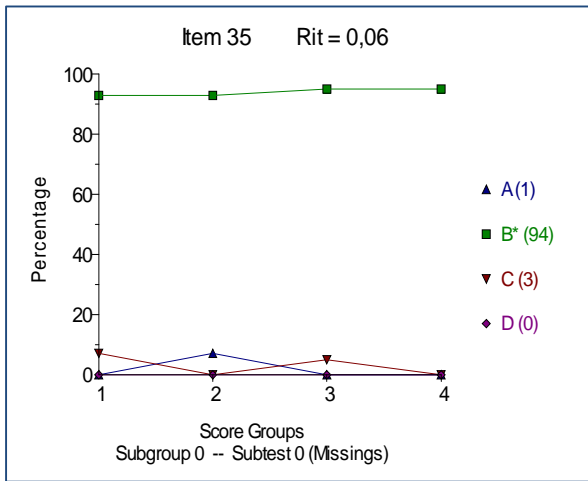
Table B.8B Poor or pathological items in *PCD-Lenguaje* Version B

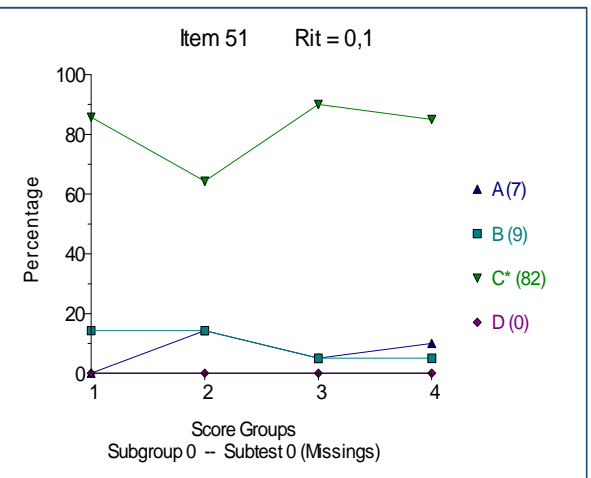
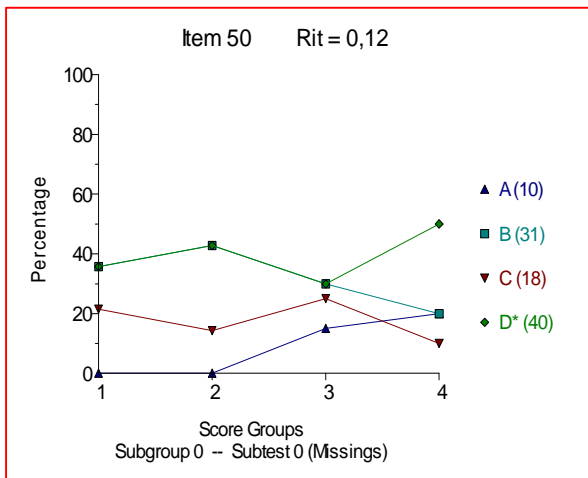
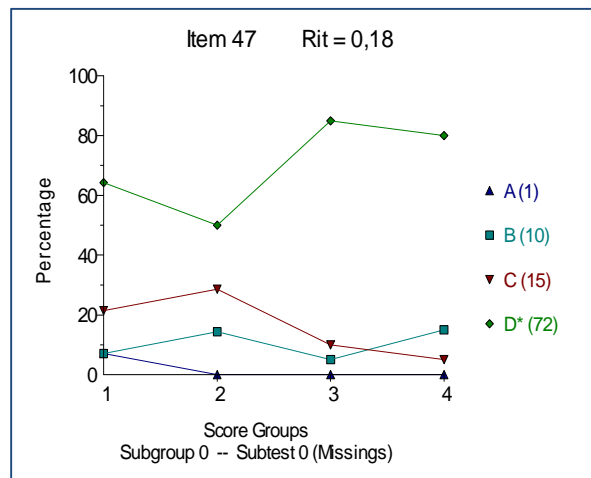
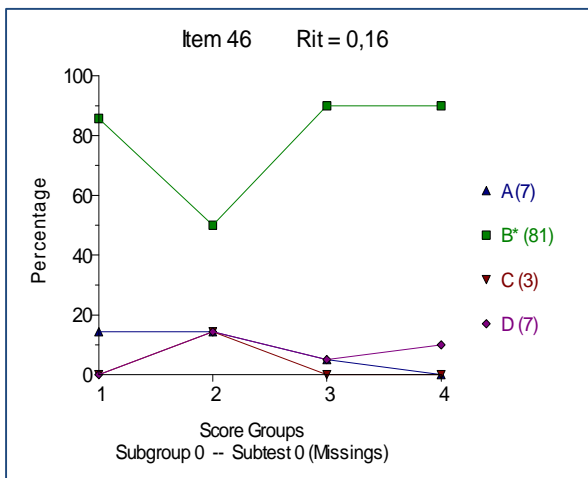
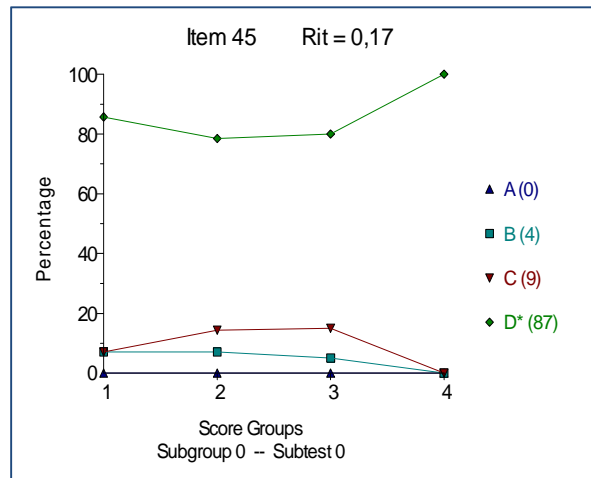
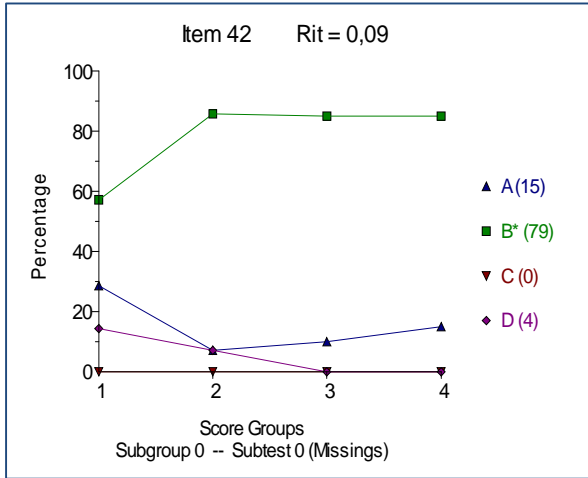
item nr.	p	Rit	Rir	Flag code ¹	Graphical analysis
1	0,99	-0,04	-0,06	ABC	There is no REAL alternative for the correct answer
2	0,94	0,01	-0,03	ABCD	This is poor/pathological because there is no REAL alternative for the correct answer and because of high guessing parameter
3	0,87	0,28	0,22	D	
6	0,53	0,18	0,1	A	There seems to be TWO correct ones (D and C) and the BEST students are messing with C
9	0,34	0,16	0,08	A	There seems to be TWO correct ones (B and C) and the BEST students are messing with C
11	0,68	0,13	0,05	A	There is no REAL alternative for the correct answer. Alternative A should be changed - no one selects it
12	0,72	0,20	0,12	BD	
16	0,84	0,20	0,14	BD	
17	0,65	0,29	0,21	D	
19	0,69	0,04	-0,05	ABC	There seems to be TWO correct ones (D and C) and the BEST students are messing with C
20	0,81	0,23	0,17	D	
24	0,51	0,18	0,09	A	There seems to be TWO correct ones (D and C) and the BEST students are messing with C. No one selects A
27	0,62	0,18	0,09	A	The BEST students are distracted by B and the POOREST ones find the correct answer too easily
30	0,75	0,08	0	ABD	There seems to be TWO correct ones (A and D) and the BEST students are messing with A. no one selects C
31	0,78	0,18	0,11	A	The BEST ones are messing with D
32	0,44	0,15	0,06	A	There seems to be TWO correct ones (B and D) and the BEST students are messing with B. (practically) no one selects A or C
34	0,93	0,01	-0,04	ABCD	There in no REAL alternative for the correct one
35	0,94	0,06	0,02	A	There in no REAL alternative for the correct one
37	0,91	0,19	0,14	A	There in no REAL alternative for the correct one
38	0,75	0,16	0,09	A	The BEST ones are messing with A and C
39	0,96	0,14	0,1	A	There in no REAL alternative for the correct one
40	0,76	0,14	0,07	A	There in no REAL alternative for the correct one and the POOREST ones are guessing the correct answer too easily
42	0,79	0,09	0,02	AB	The BEST ones are messing with A
45	0,87	0,17	0,12	A	There in no REAL alternative for the correct one and the POOREST ones are guessing the correct answer too easily
46	0,81	0,16	0,09	ABD	There in no REAL alternative for the correct one and the POOREST ones are guessing the correct answer too easily
47	0,72	0,18	0,1	A	The BEST ones are messing with B and the POOREST ones are guessing the correct answer too easily
50	0,4	0,12	0,03	ABD	There in NO correct answer
51	0,82	0,1	0,03	AB	There in no REAL alternative for the correct one and the POOREST ones are guessing the correct answer too easily
53	0,85	-0,02	-0,08	ABCD	The BEST ones are messing with D and the POOREST ones know the correct answer too easily
56	0,79	0,14	0,07	ABD	There in no REAL alternative for the correct one and the BEST ones are messing with D

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high









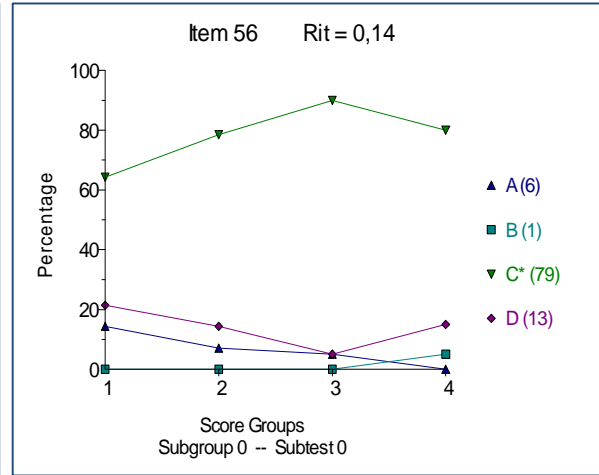
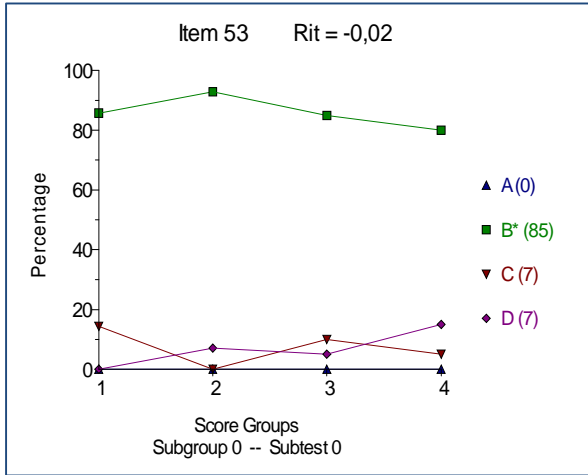
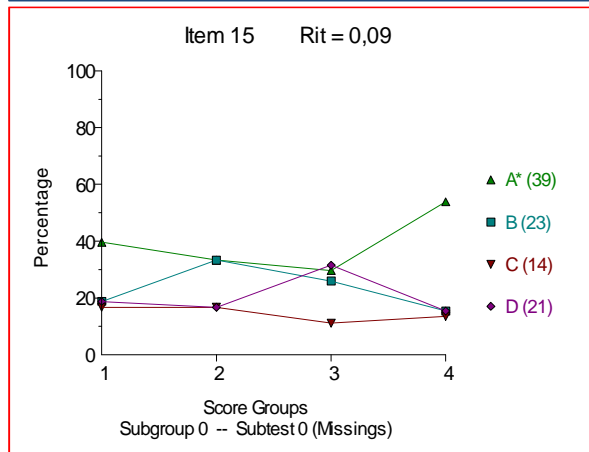
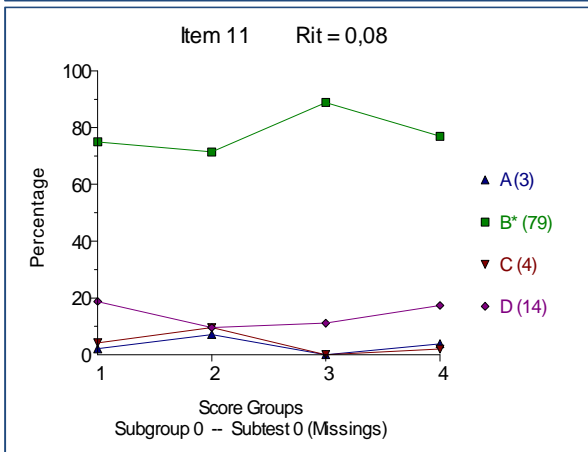
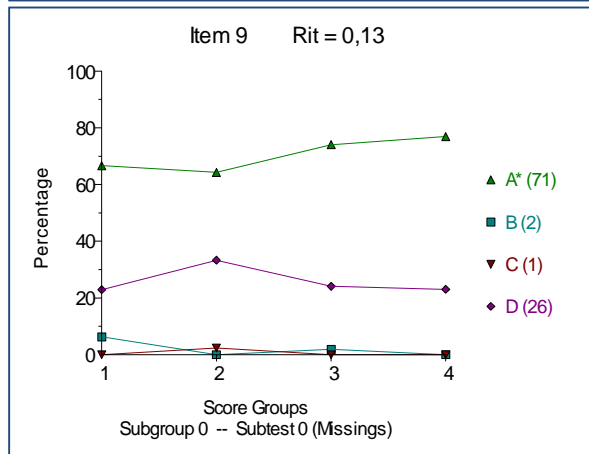
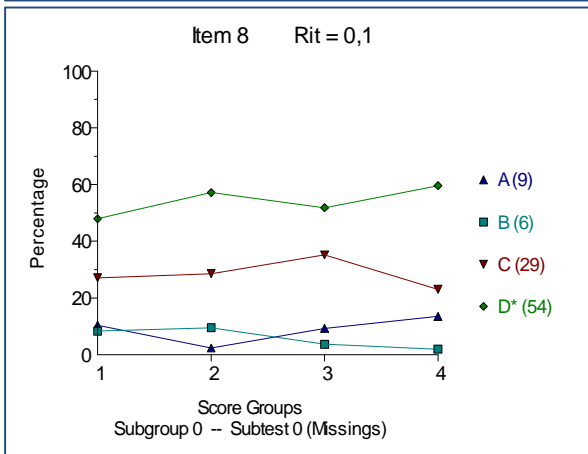
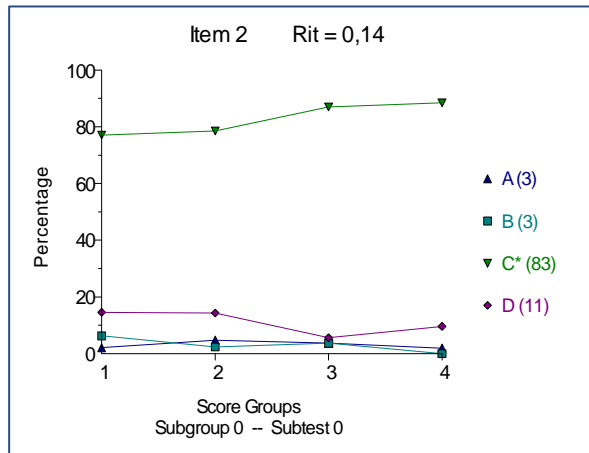
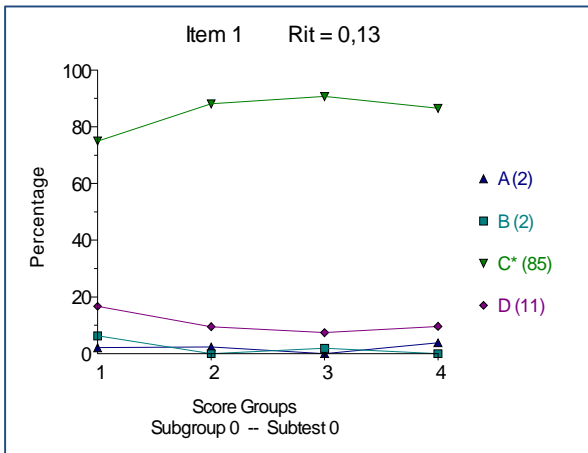


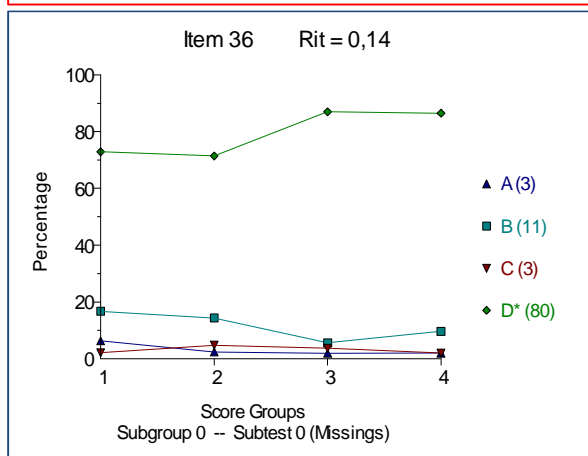
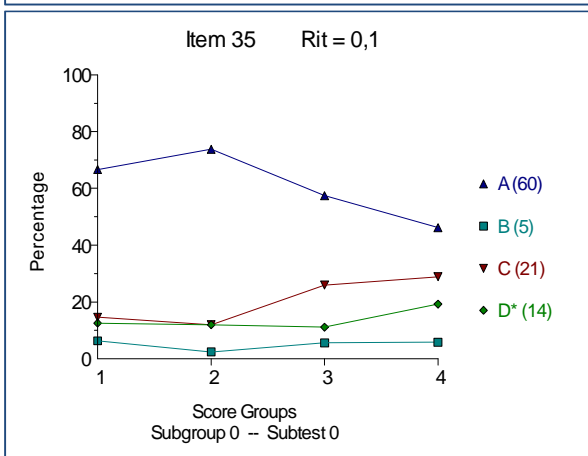
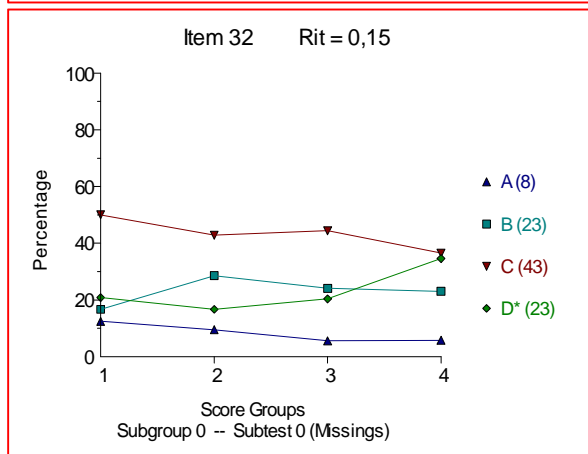
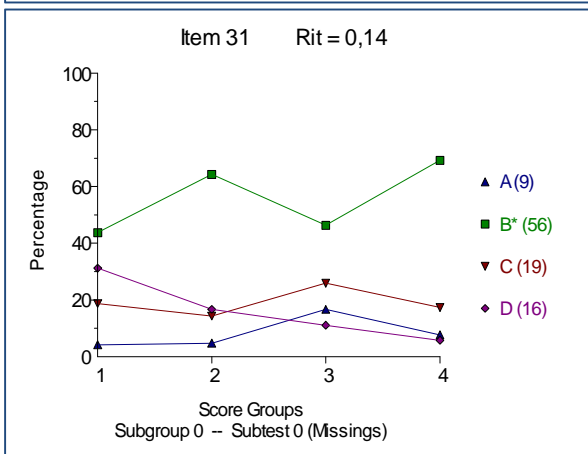
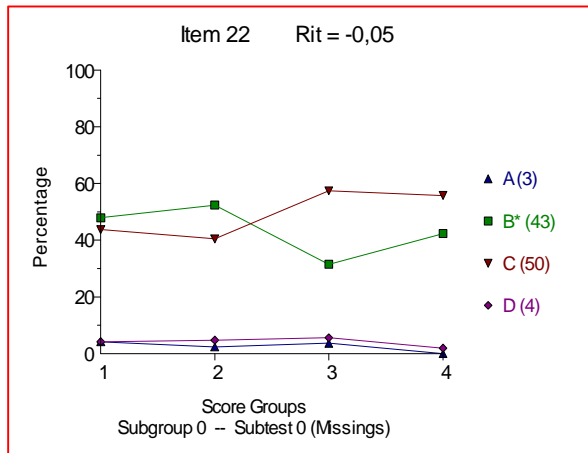
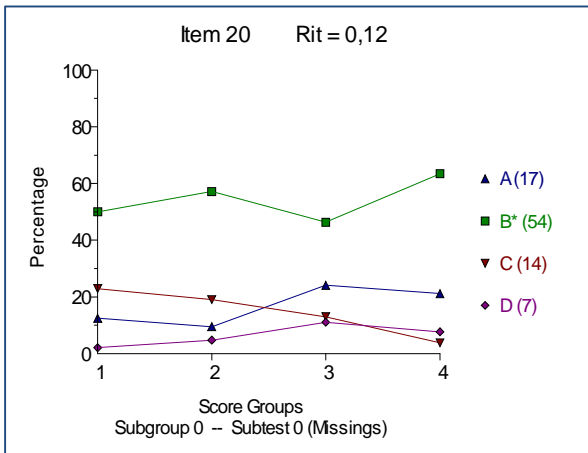
Table B.9A Poor or pathological items in PCD-Parvularia Version A

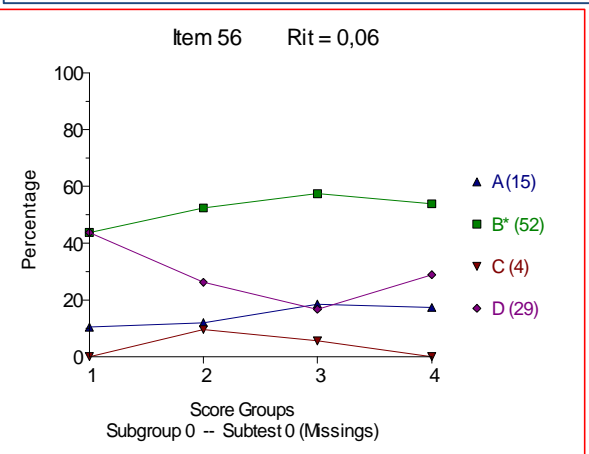
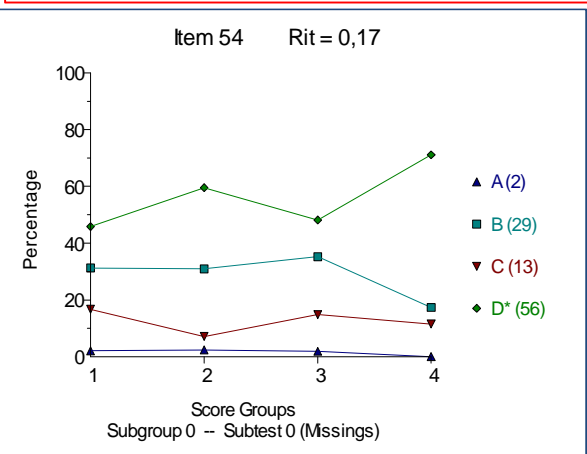
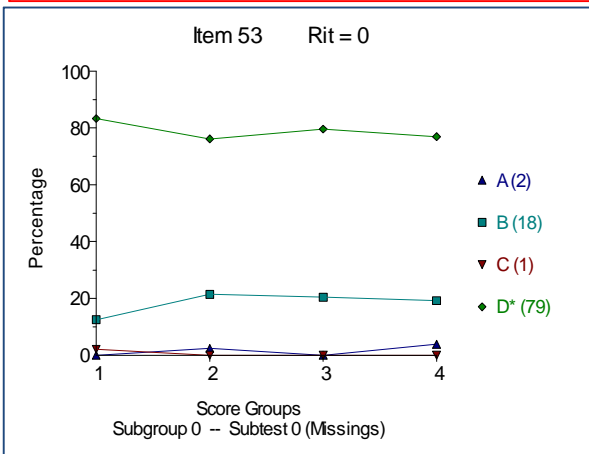
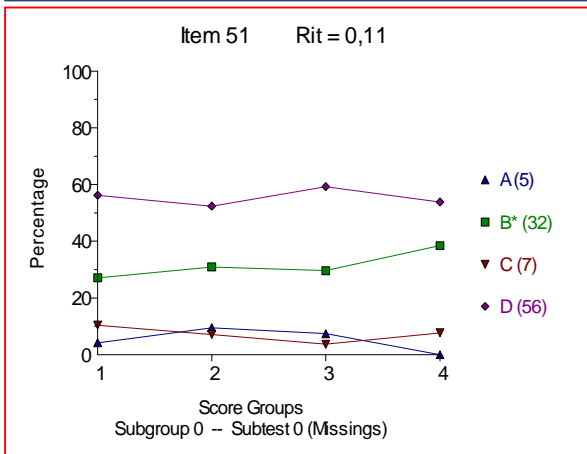
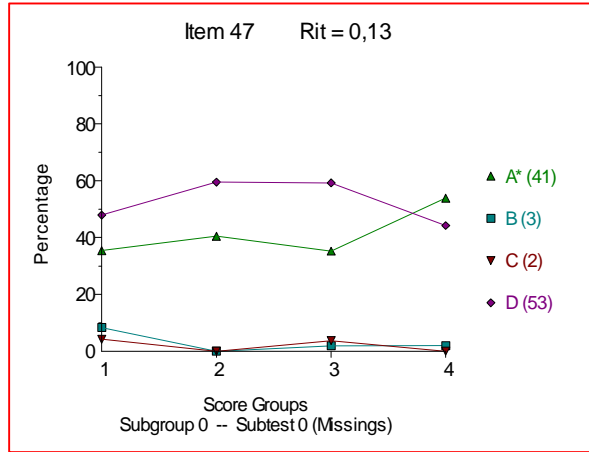
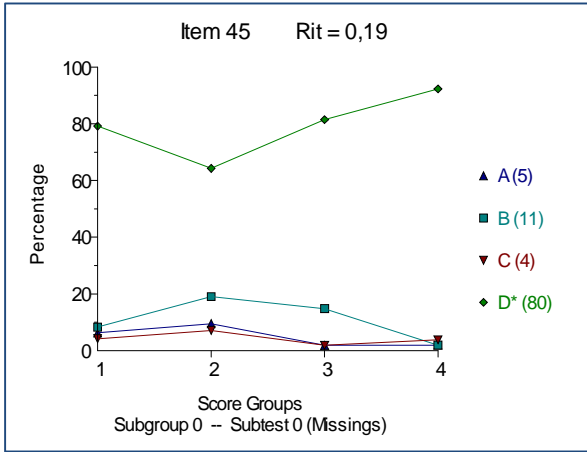
item nr.	p	Rit	Rir	Flag code ¹	Graphical analysis
1	0,85	0,13	0,08	A	There is no REAL alternative for the correct answer
2	0,83	0,14	0,08	A	There is no REAL alternative for the correct answer
5	0,7	0,14	0,07	A	There is no REAL alternative for the correct answer and the weakest students find the correct alternative too easily
8	0,54	0,1	0,02	AB	There seems to be TWO alternatives for the correct answer (D and C)
9	0,71	0,13	0,06	A	There seems to be TWO alternatives for the correct answer (D and A). Also, the weakest students find the correct alternative too easily
11	0,79	0,08	0,02	AB	The BEST ones are distracted to (D). check the key!
15	0,39	0,09	0,01	AB	The WEAKEST students guess too easily the correct answer
20	0,54	0,12	0,04	ABD	The POOREST guess the correct alternative and the many of the BEST ones are distracted to A
22	0,43	-0,05	-0,13	ABCD	This is pathological because there seems to be TWO alternatives for the correct answer (B and C). Check the key! (C would be more plausible)
31	0,56	0,14	0,06	AB	There seems to be specific knowledge in group 2 and the poorest find the correct answer too easily
32	0,23	0,15	0,08	AB	This is pathological because there seems to be TWO alternatives for the correct answer (D and C). OR there is NO correct answer!
33	0,29	0,24	0,17	BD	
35	0,14	0,1	0,04	ABD	This is pathological because the BEST students do not know the correct answer. This is just POOR item.
36	0,8	0,14	0,08	A	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
45	0,8	0,19	0,13	A	There is no REAL alternative for the correct answer and because the weakest students find the correct alternative too easily
47	0,41	0,13	0,05	A	This is pathological because there seems to be TWO alternatives for the correct answer (A and D).
51	0,32	0,11	0,04	AB	This is pathological because there seems to be TWO alternatives for the correct answer (A and D). The BEST ones are messing with D
53	0,79	0	-0,07	ABCD	This is pathological because the BEST students do not know the correct answer and because the weakest students find the correct alternative too easily. This is just POOR item.
54	0,56	0,17	0,09	A	The weakest students find the correct alternative too easily
55	0,64	0,26	0,19	D	
56	0,52	0,06	-0,02	ABCD	This is pathological because the BEST students do not know the correct answer. They mess with D and A. This is just poor item.
57	0,55	0,18	0,1	A	This is pathological because the BEST students do not know the correct answer. They are messing with B. Also

60	0,17	0,17	0,11	ABD	the weakest students find the correct alternative too easily This is pathological. Seems that there are TWO correct answers (C and D). Of these the C would be more probable. Check the key!
----	------	------	------	-----	---

-
- 1) A: $R_{it} < 0.20$ item-total correlation is low, B: $R_{ar} \geq R_{ir}$ a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: $R_{ir} \leq 0$ the correct alternative does not correlate or even correlates negatively with the test's rest score, D: $R_{ar} \geq 10$ a distracter - test score correlation is suspiciously high







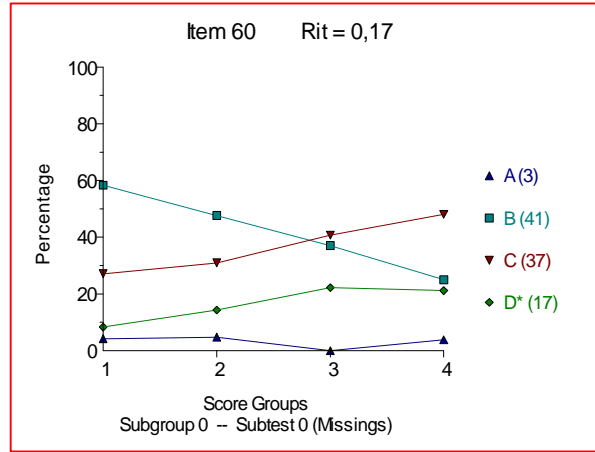
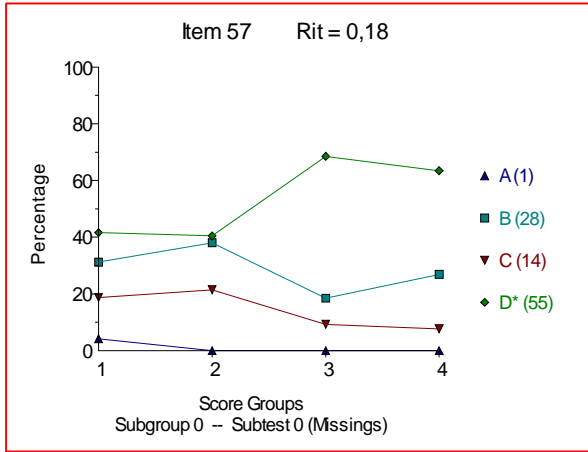
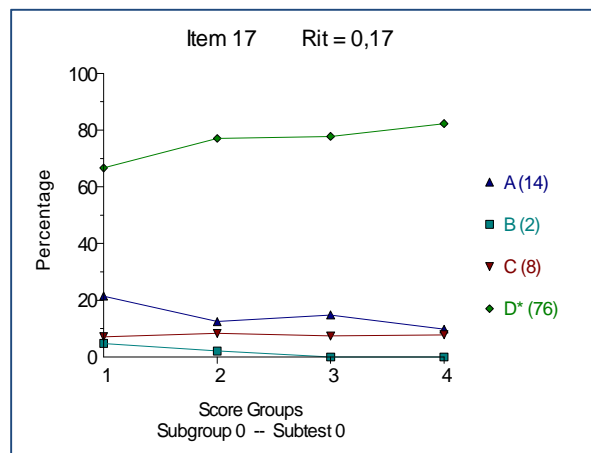
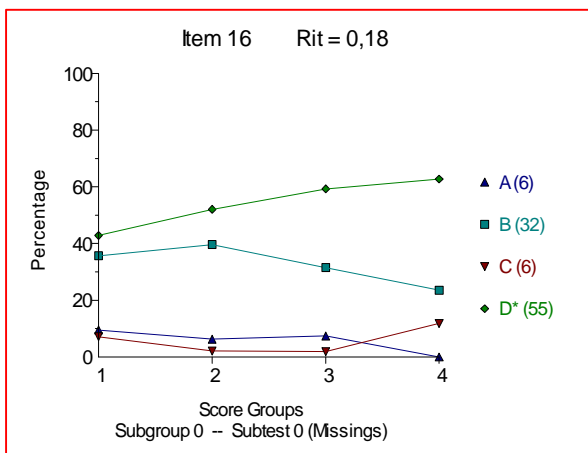
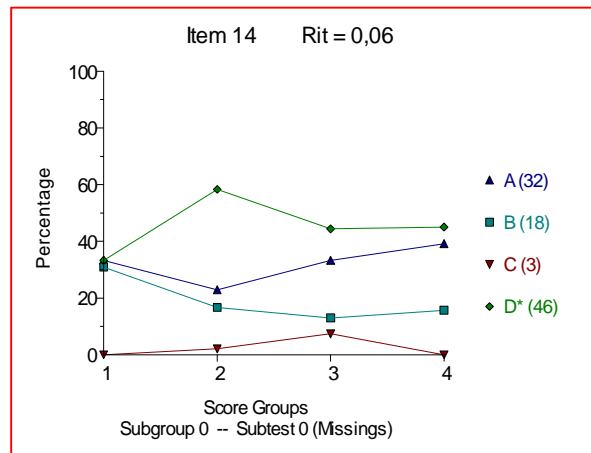
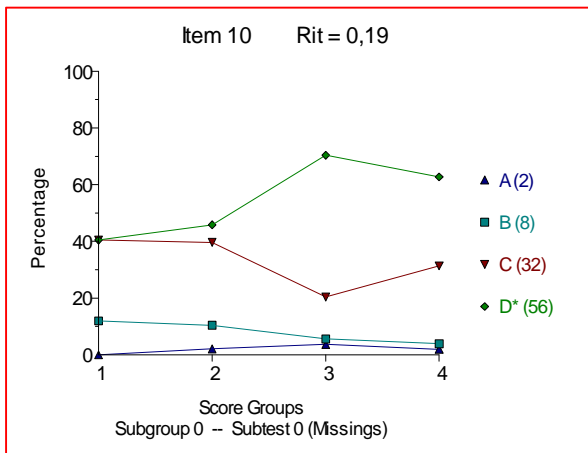
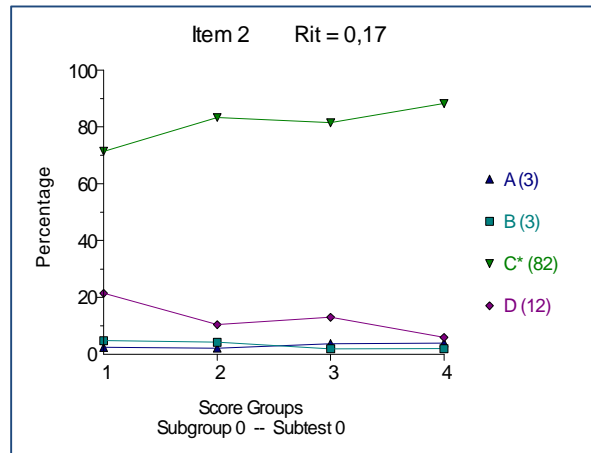
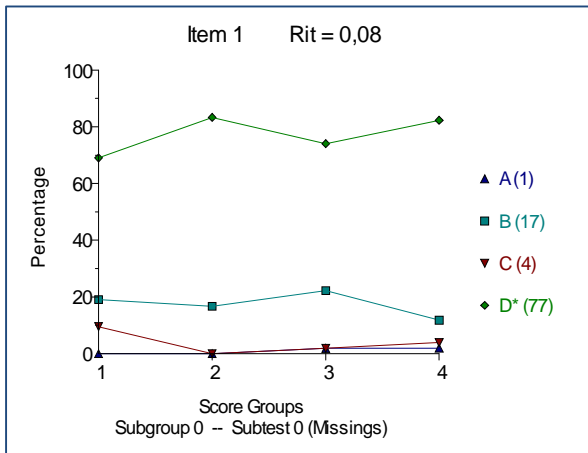
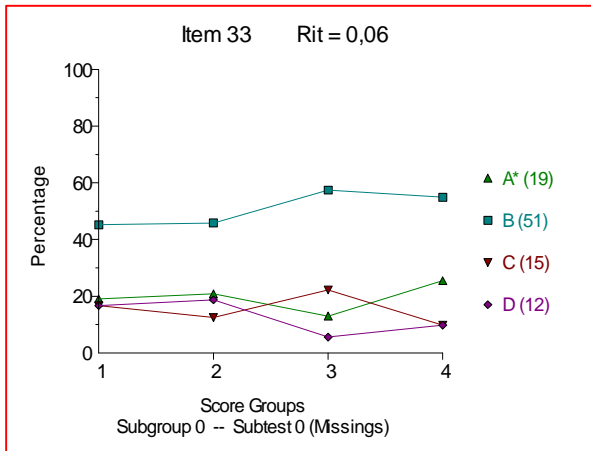
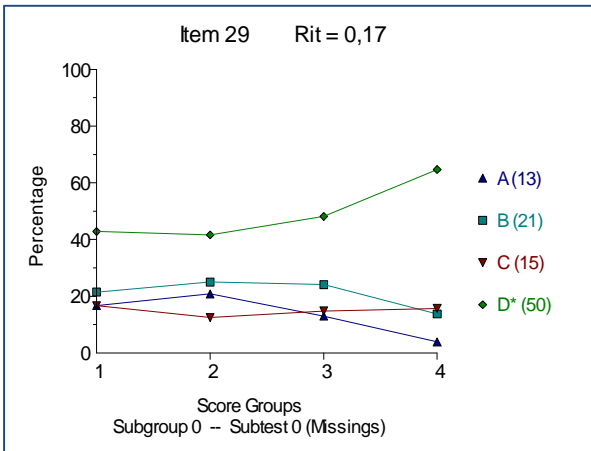
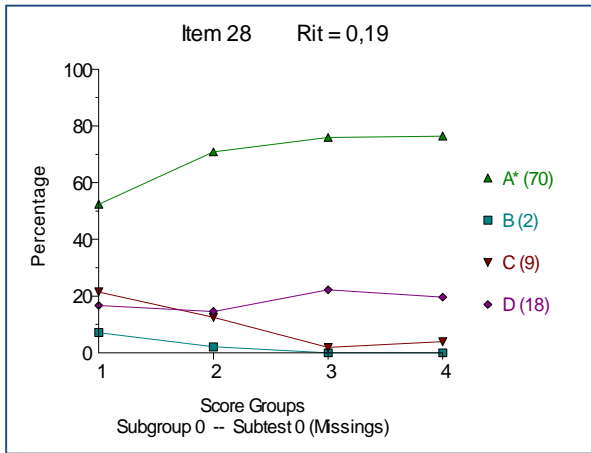
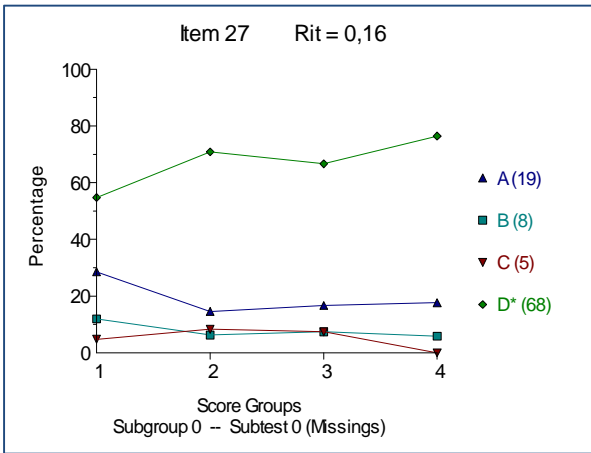
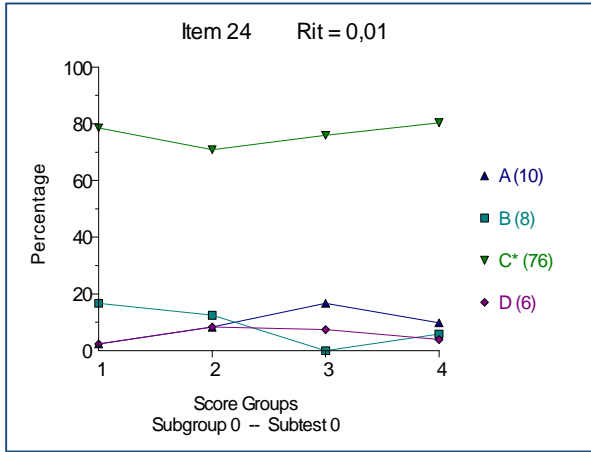
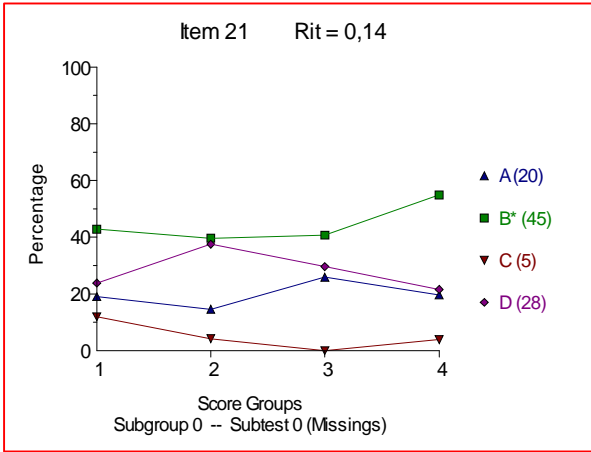


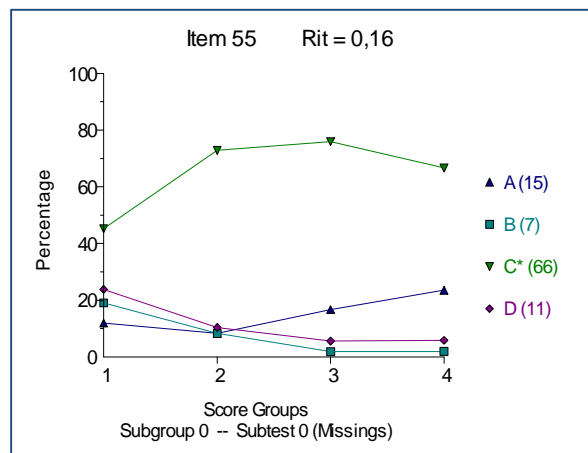
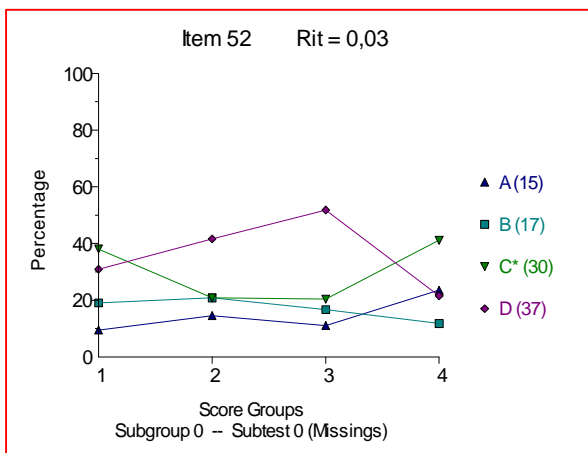
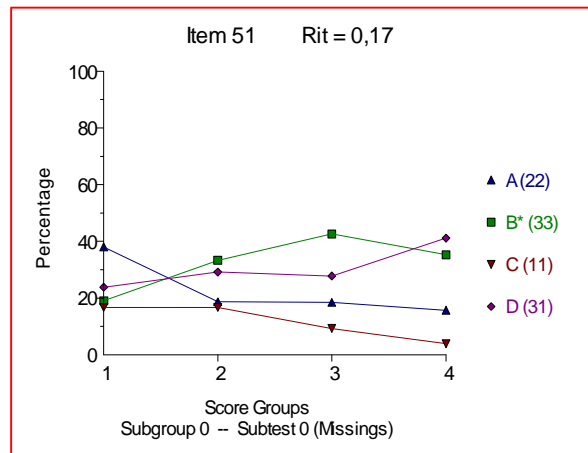
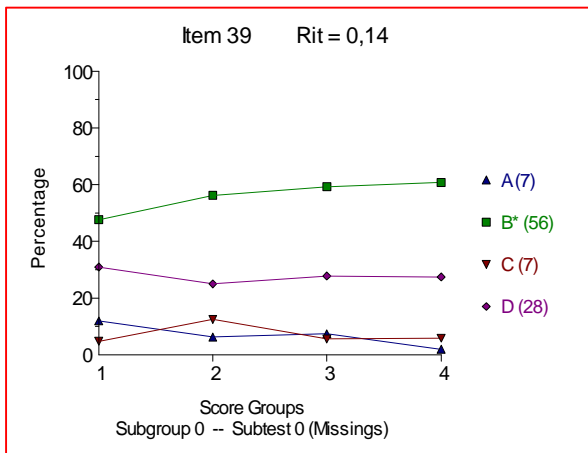
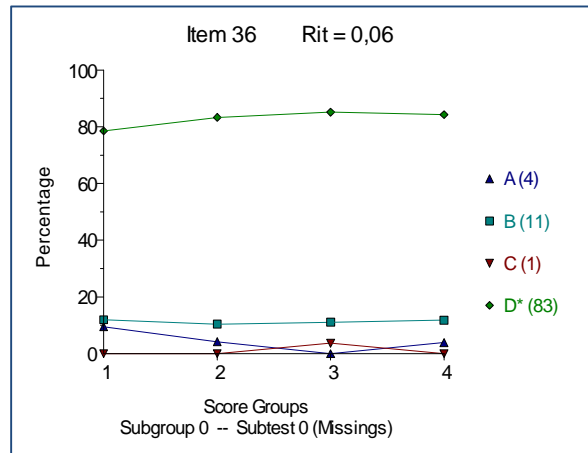
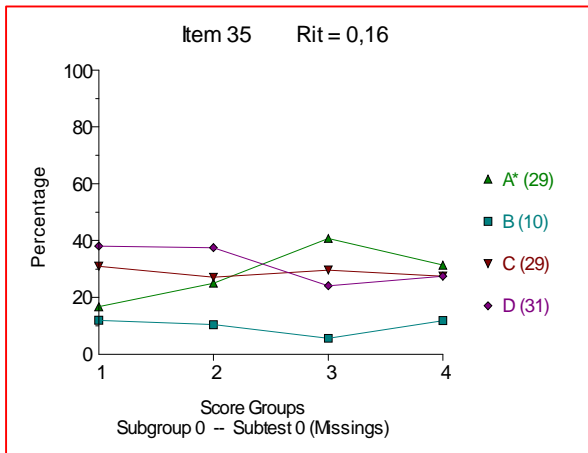
Table B.9B Poor or pathological items in PCD-Parvularia Version B

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
1	0,77	0,08	0,02	ABD	There is no REAL alternative for the correct answer and because the WEAKEST ones know the correct answer too easily
2	0,82	0,17	0,11	A	There is no REAL alternative for the correct answer
10	0,56	0,19	0,12	A	There are TWO correct alternatives (C and D) and the BEST students are messing with C
11	0,64	0,23	0,16	D	
13	0,51	0,21	0,14	D	
14	0,46	0,06	-0,02	ABC	There are TWO correct answers (D and A) or NO correct answer
16	0,55	0,18	0,1	A	There are TWO correct alternatives (D and B) and the BEST students are messing with B
17	0,76	0,17	0,1	A	There is no REAL alternative for the correct answer and the WEAKEST students find the correct alternative too easily
21	0,45	0,14	0,07	A	There are THREE correct answers or NO correct answer
24	0,76	0,01	-0,06	ABCD	There is no REAL alternative for the correct answer and the WEAKEST students find the correct alternative too easily
27	0,68	0,16	0,09	A	The BEST students are messing with A
28	0,7	0,19	0,12	A	The BEST students are messing with D
29	0,5	0,17	0,1	A	The weakest students find the correct alternative too easily
30	0,46	0,25	0,17	D	
33	0,19	0,06	0	ABCD	This is pathological because the BEST students are messing with B and the WEAKEST ones find the correct one too easily Check the key! B could be a correct one.
35	0,29	0,16	0,09	A	There are THREE correct answers
36	0,83	0,06	0	AB	There is no REAL alternative for the correct alternative
39	0,56	0,14	0,06	A	There seems to be TWO correct answers
51	0,33	0,17	0,09	ABD	There seems to be TWO correct answers BEST ones are messing with D. Check the key!
52	0,3	0,03	-0,04	ABCD	This is pathological because the WEAKEST students find the correct answer too easily. Check the key! D?
55	0,66	0,16	0,09	ABD	The BEST ones are messing with A. Check the key! A could be correct!
56	0,59	0,15	0,07	A	The BEST ones do not find the correct answer
59	0,32	0,18	0,11	ABD	There are THREE correct answers
60	0,27	0,11	0,04	AB	There is NO correct answer

1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high







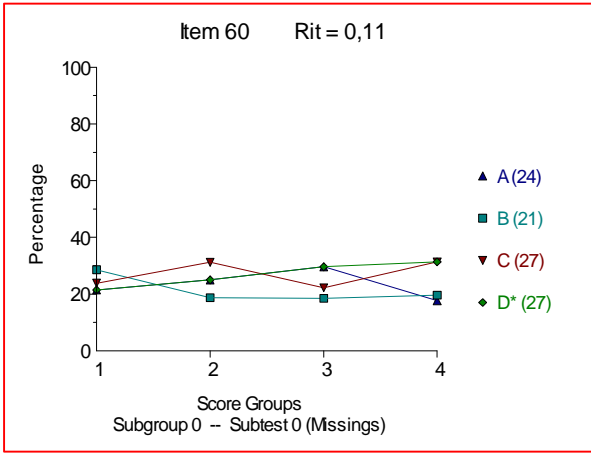
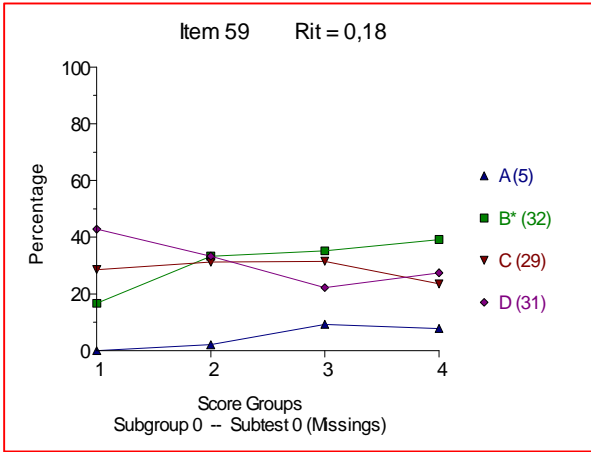
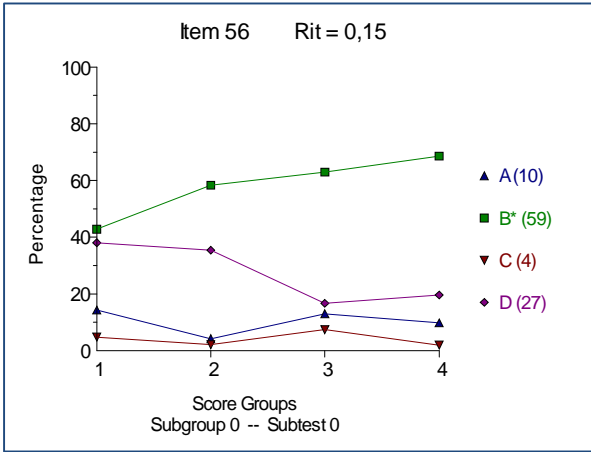
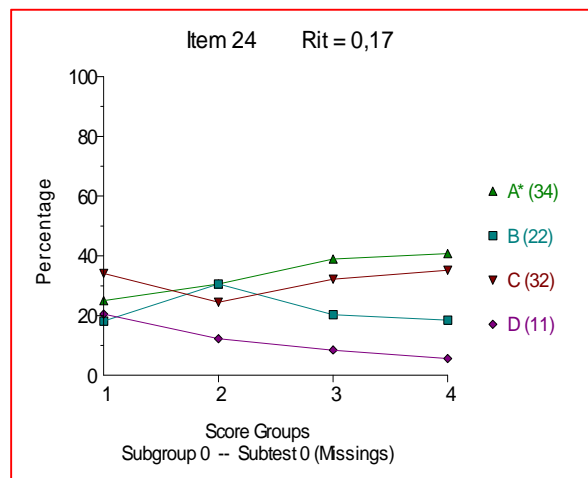
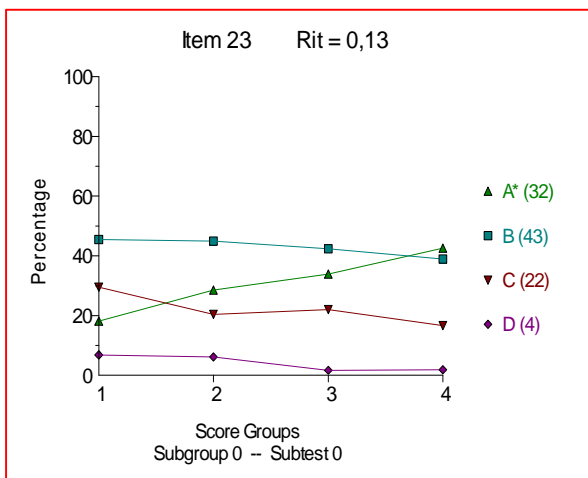
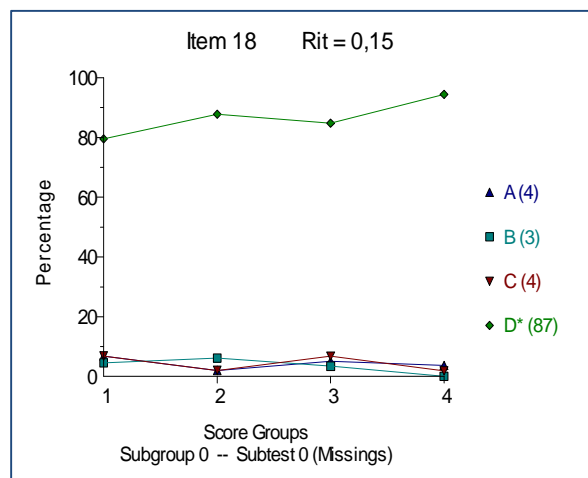
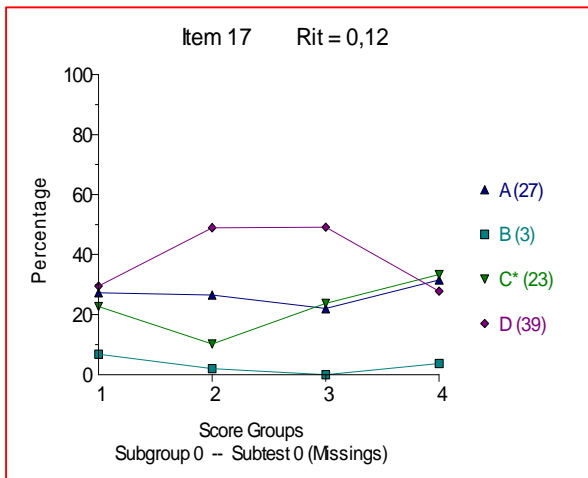
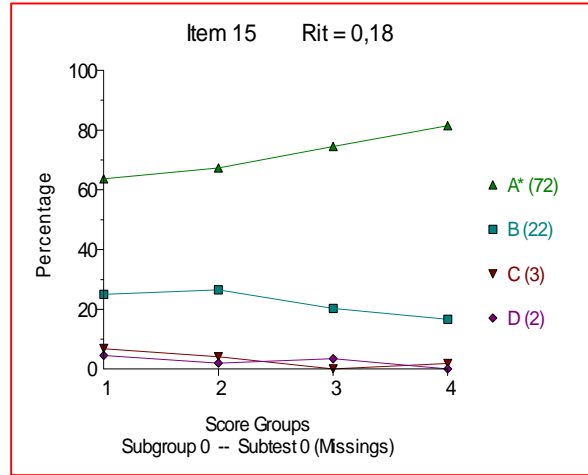
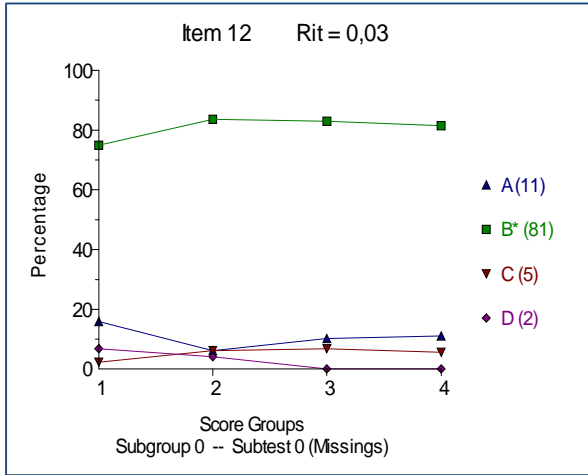


Table B.10A Poor or pathological items in PCP-Parvularia Version A

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
12	0,81	0,03	-0,04	ABC	The BEST ones do not find the correct answer and WEAKEST ones find the correct answer too easily
13	0,51	0,03	-0,06	ABCD	There are TWO correct answers (D and B)
15	0,72	0,18	0,1	A	The WEAKEST ones find the correct answer too easily
17	0,23	0,12	0,04	AB	There seems to be THREE correct answers and the BEST ones do not find the correct one
18	0,87	0,15	0,09	A	There is no REAL alternative for the correct answer
20	0,66	0,2	0,12	BD	
23	0,32	0,13	0,05	A	There seems to be TWO correct answers (A and B) and the BEST ones do not find the correct one. They mess with B
24	0,34	0,17	0,09	A	There seems to be TWO correct answers (A and C) and the BEST ones do not find the correct one. They mess with C
26	0,53	0,17	0,08	A	The WEAKEST ones find the correct alternative too easily
28	0,51	0,19	0,11	A	There seems to be TWO correct answers (D and B) and the BEST ones do not find the correct one. They mess with B
31	0,25	0,18	0,1	A	There seems to be TWO correct answers (C and B) and the BEST ones do not find the correct one. They mess with C
34	0,51	0,16	0,07	A	The WEAKEST ones find the correct alternative too easily
42	0,46	0,16	0,08	A	The WEAKEST ones find the correct alternative too easily
44	0,4	0,08	0	ABC	This is pathological because the WEAKEST find the correct alternative too easily and because there seems to be TWO correct answers (C and D)
48	0,78	0,05	-0,02	ABC	The WEAKEST find the correct alternative too easily and the BEST ones do not find the correct one. They mess with B
50	0,87	0,08	0,02	ABD	There is no REAL alternative for the correct answer

1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high



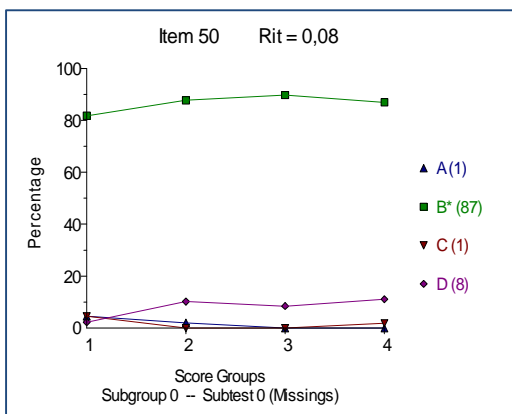
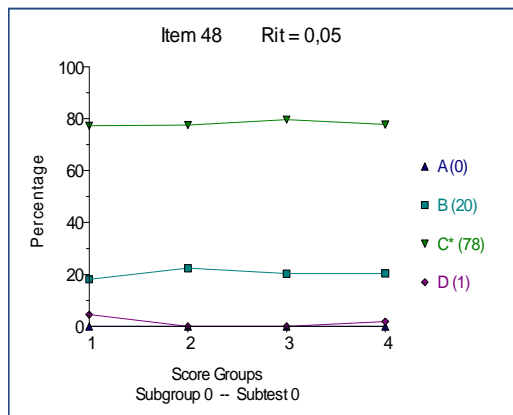
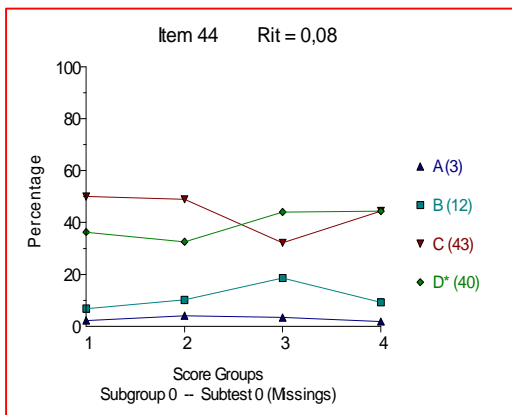
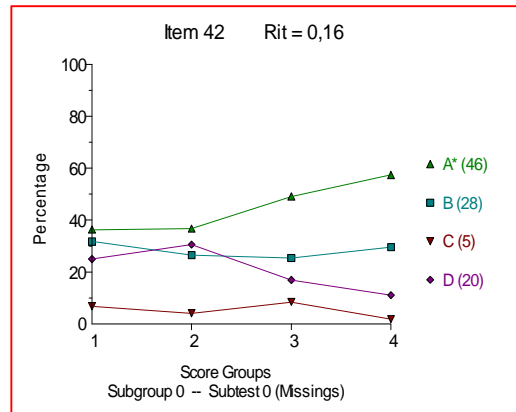
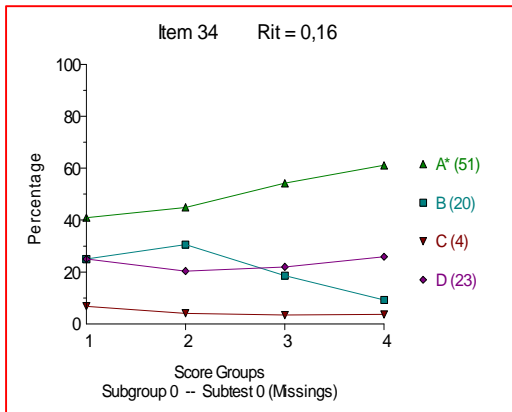
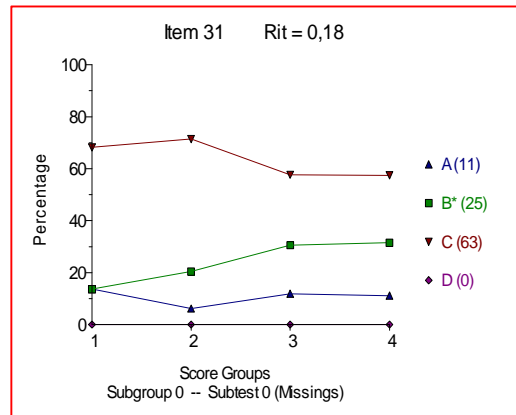
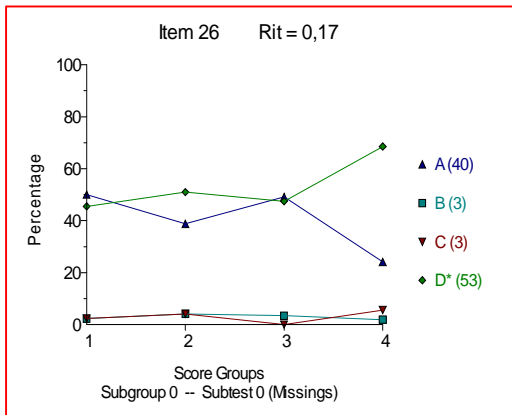
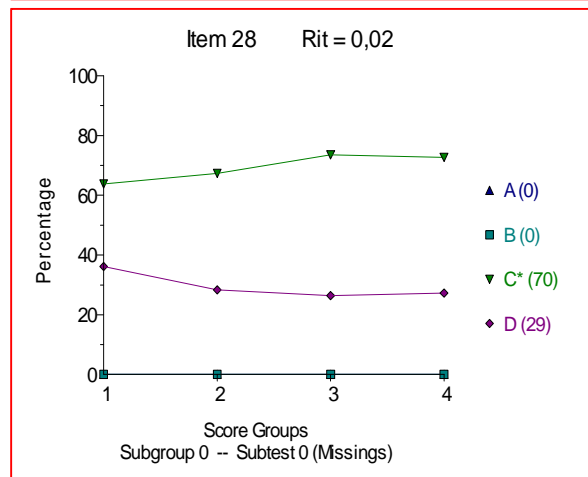
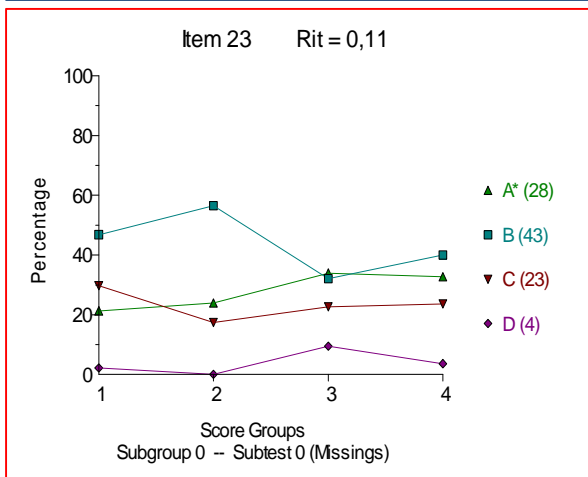
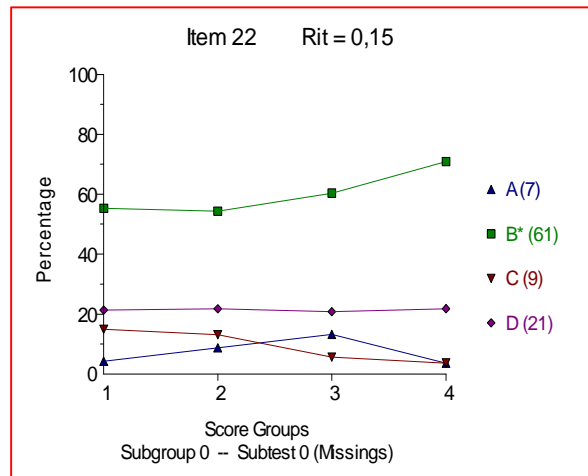
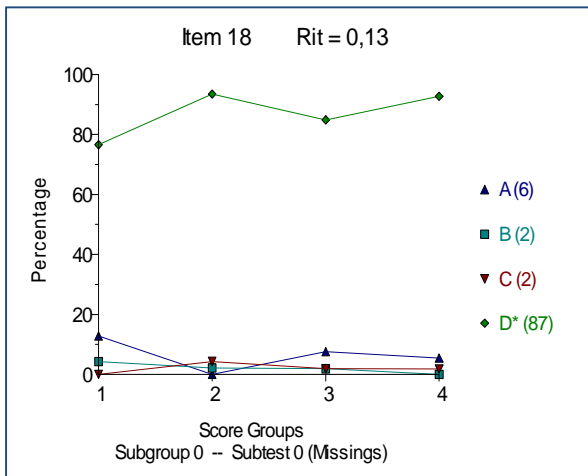
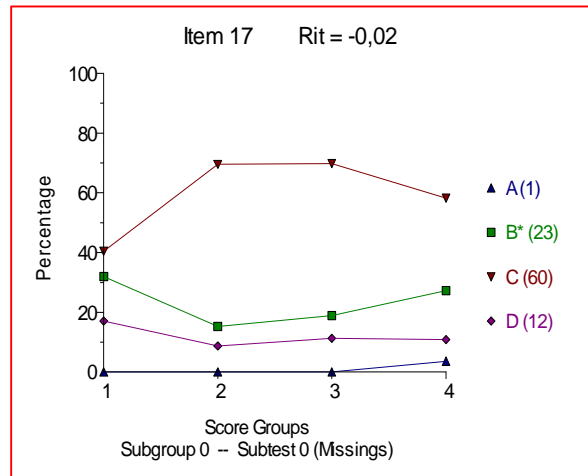
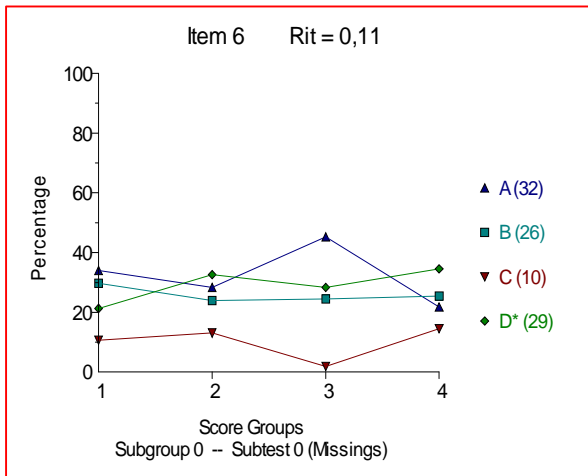


Table B.10B Poor or pathological items in PCP-Parvularia Version B

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
6	0,29	0,11	0,03	AB	There is no correct answer
8	0,34	0,2	0,12	BD	
17	0,23	-0,02	-0,09	ABCD	The WEAKEST ones find the correct answer too easily the BEST ones does not find the correct one. They are messing A. Check the key. Poor item.
18	0,87	0,13	0,08	A	There is no REAL alternative for the correct answer
22	0,61	0,15	0,07	A	The WEAKEST ones find the correct answer too easily and the BEST ones are distracted by D.
23	0,28	0,11	0,03	AB	There is no correct answer
28	0,7	0,02	-0,06	ABC	There seems to be TWO correct answers (C and D) and the BEST ones do not find the correct one. They mess with D and the WEAKEST ones find the correct answer too easily
30	0,2	0,1	0,03	AB	There seems to be TWO correct answers (C and D) and the BEST ones do not find the correct one. They mess with D and the WEAKEST ones find the correct answer too easily
32	0,27	0,17	0,1	A	There seems to be TWO correct answers (B and C) and the BEST ones do not find the correct one. They mess with A
35	0,51	0,16	0,08	A	The WEAKEST ones find the correct answer too easily
46	0,78	0,16	0,09	A	The WEAKEST ones find the correct answer too easily
48	0,51	0,15	0,07	ABD	The WEAKEST ones find the correct answer too easily

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high



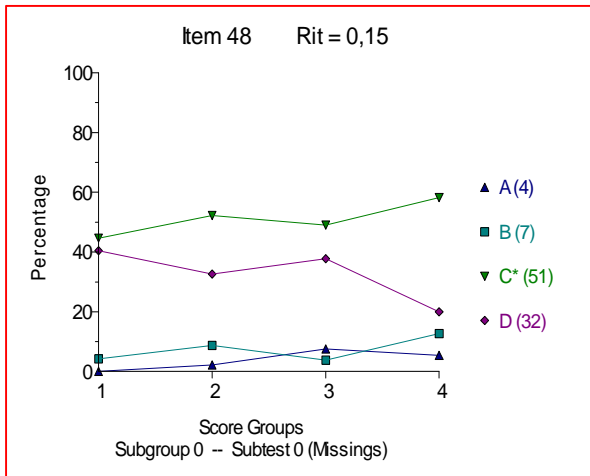
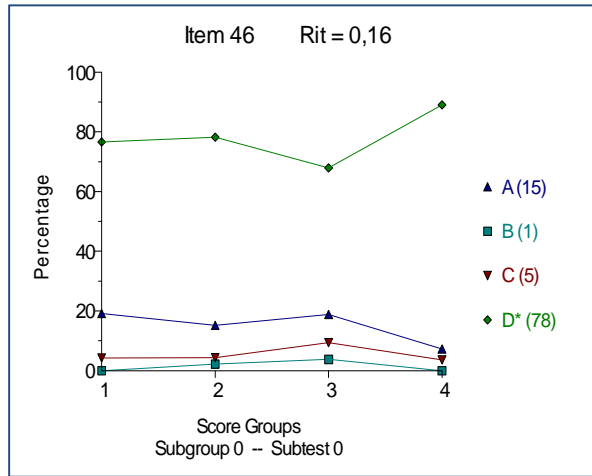
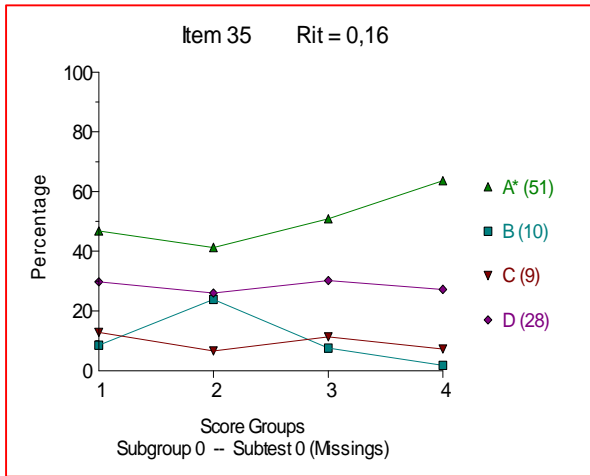
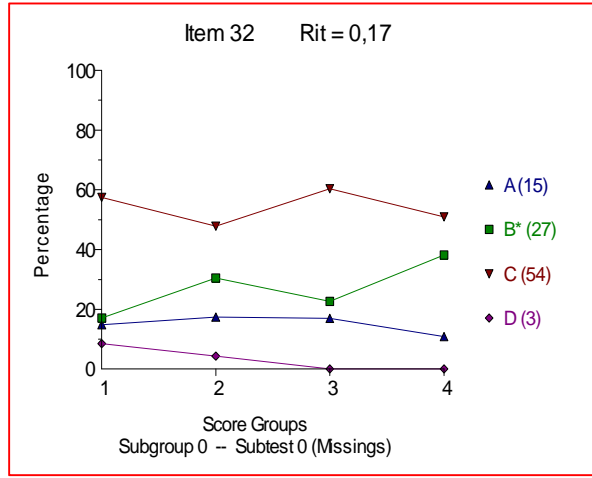
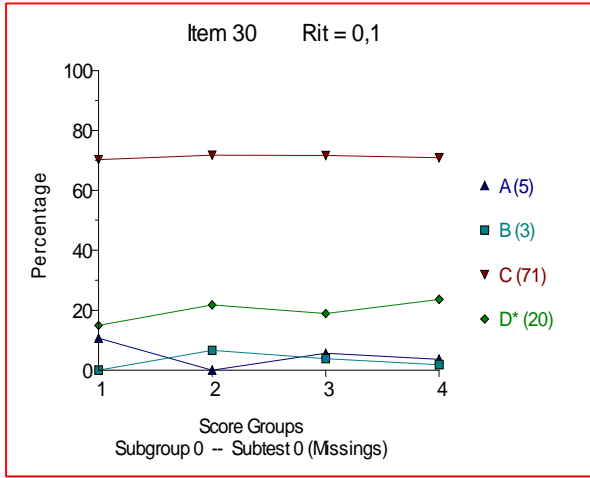
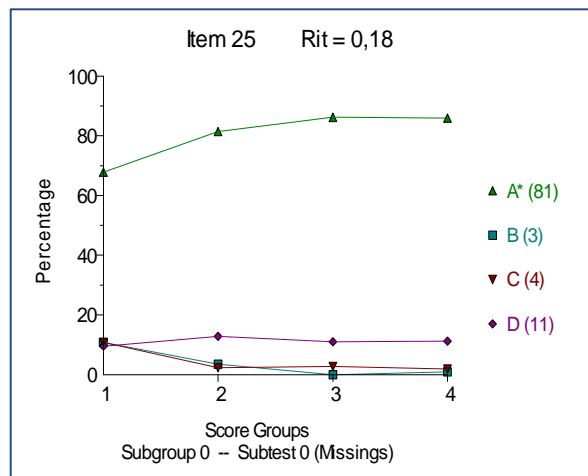
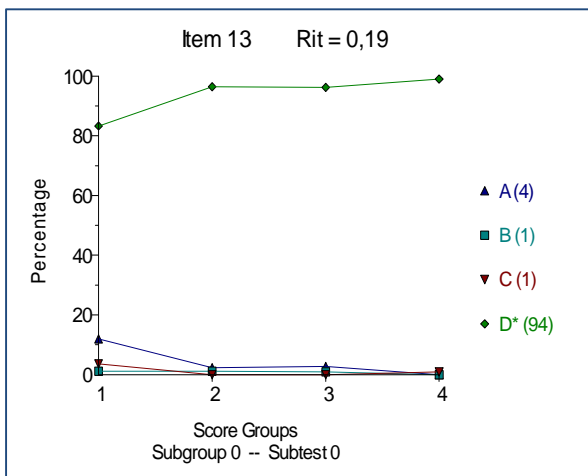
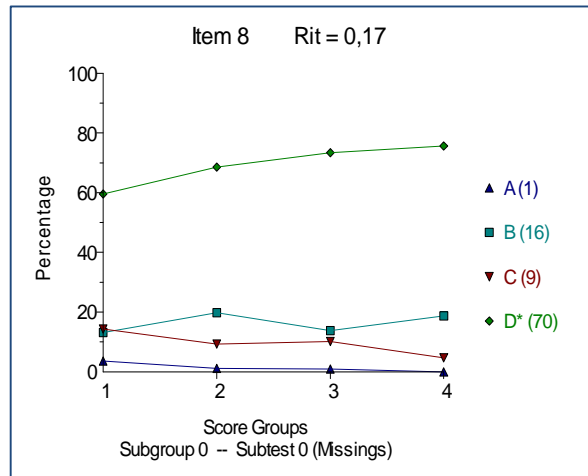
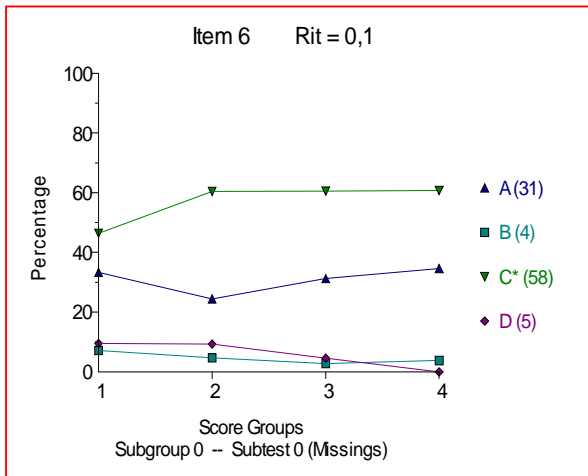
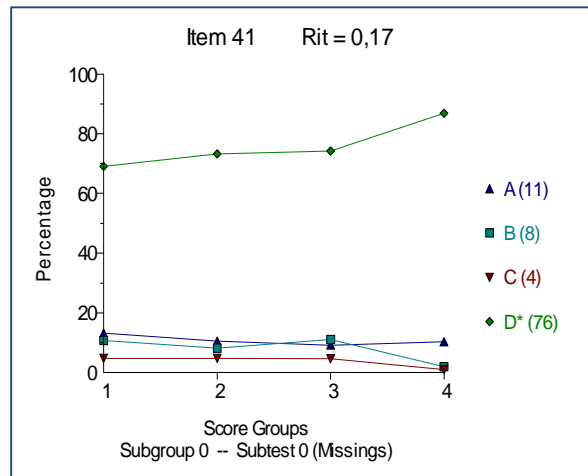
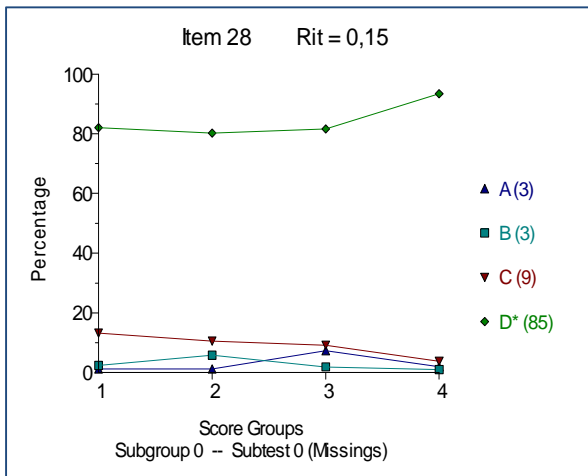
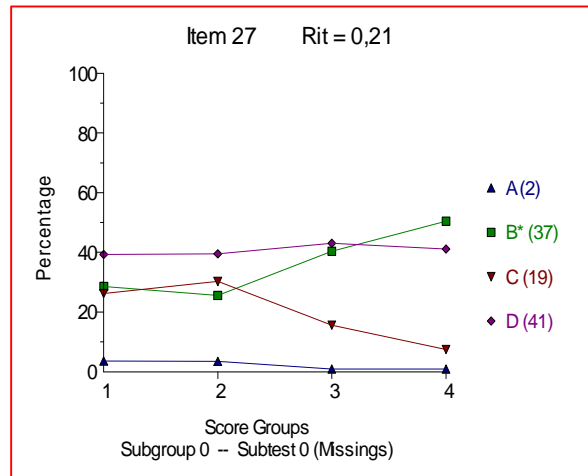
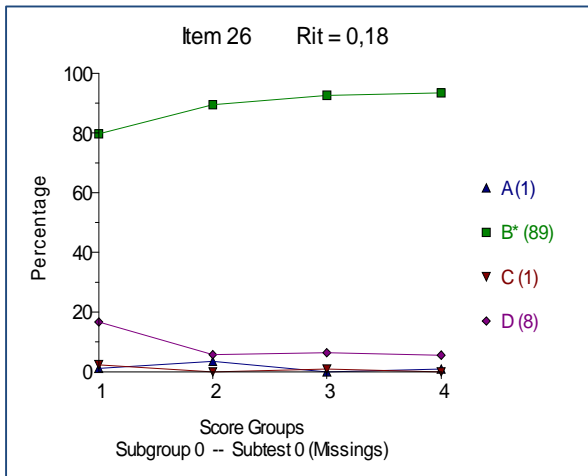


Table B.11A Poor or pathological items in PCP-Media Version A

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
6	0,58	0,10	0,01	ABD	The BEST ones do not find the correct answer; they are distracted by A. Check the key. Is A the real key?
8	0,7	0,17	0,09	A	There is no REAL alternatives for the correct answer and the WEAKEST ones find the correct answer too easily
13	0,94	0,19	0,15	A	There is no REAL alternative for the correct answer
25	0,81	0,18	0,11	A	There is no REAL alternative for the correct answer
26	0,89	0,18	0,13	A	There is no REAL alternative for the correct answer and the WEAKEST students find the correct alternative too easily
27	0,37	21	13		The BEST students seems to be distracted to D. There seems to be TWO correct answers (B and D)
28	0,85	0,15	0,08	A	The POOREST find the correct alternative too easily. There is no REAL alternative for the correct answer
35	0,62	0,18	0,10	A	no problem
41	0,76	0,17	0,10	A	There is no REAL alternatives for the correct answer and the WEAKEST ones find the correct answer too easily
46	0,79	0,16	0,09	A	The WEAKEST ones find the correct alternative too easily
47	0,66	0,15	0,07	A	The WEAKEST ones find the correct alternative too easily
50	0,5	0,18	0,09	A	The WEAKEST ones find the correct alternative too easily

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high





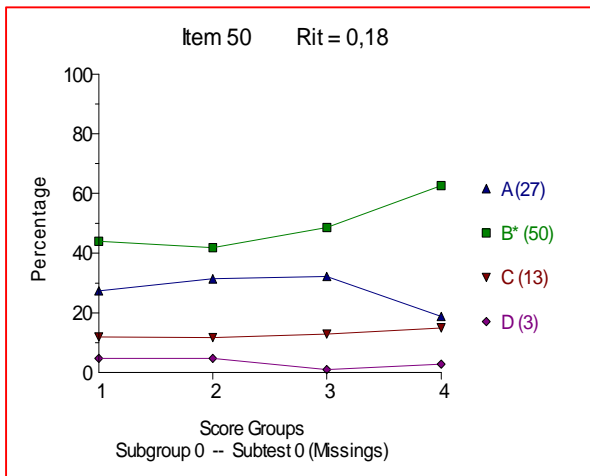
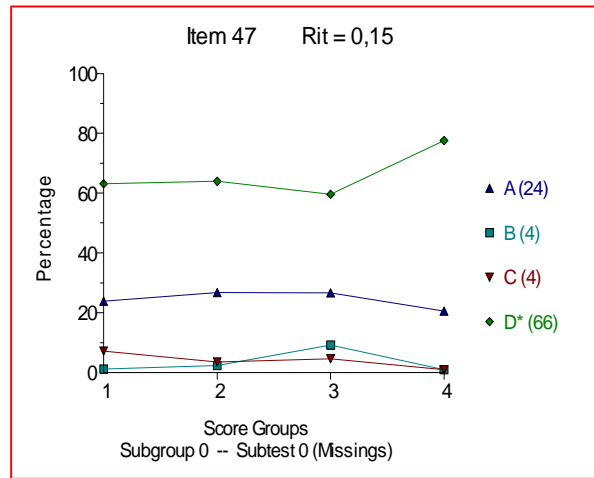
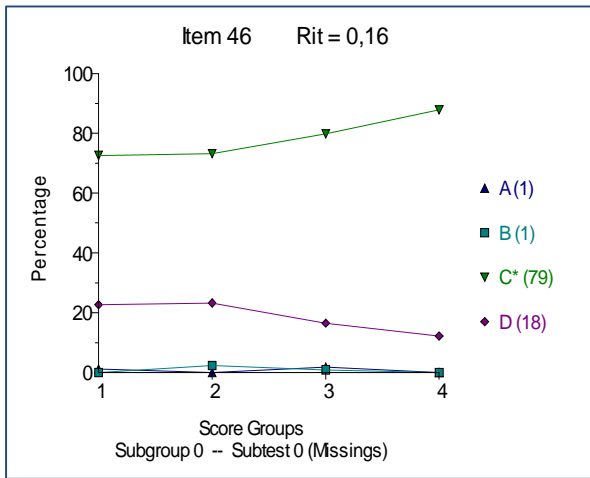
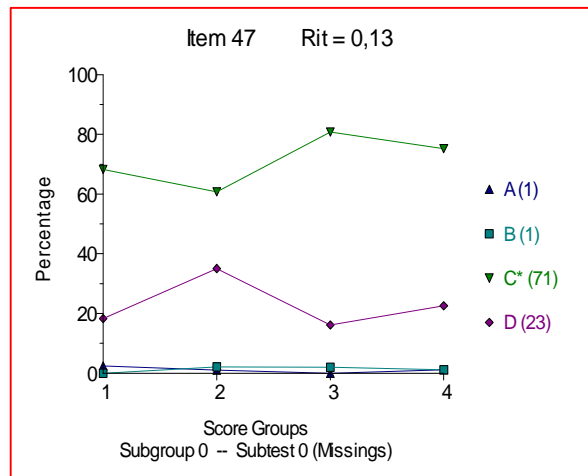
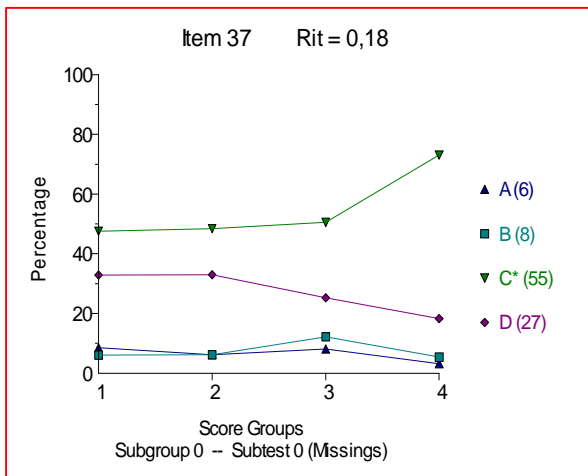
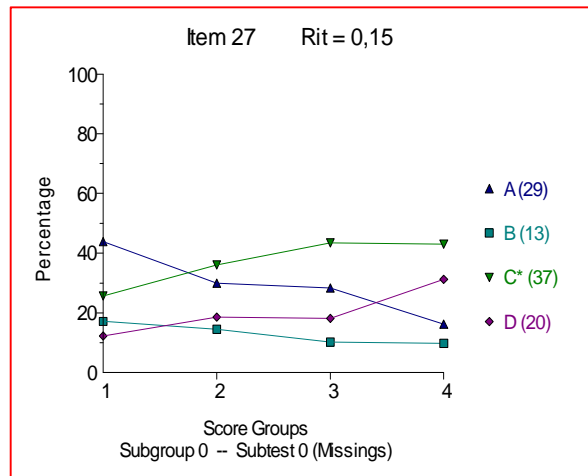
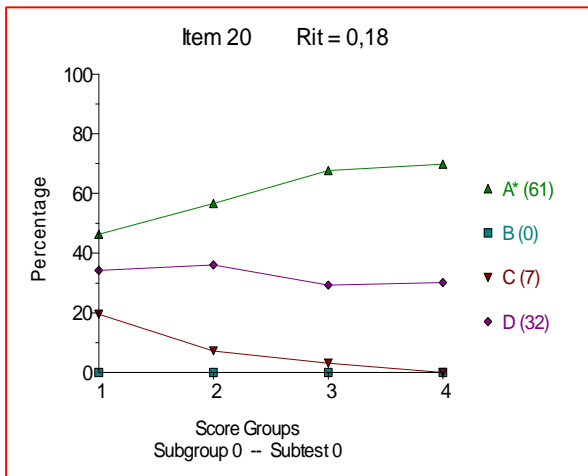
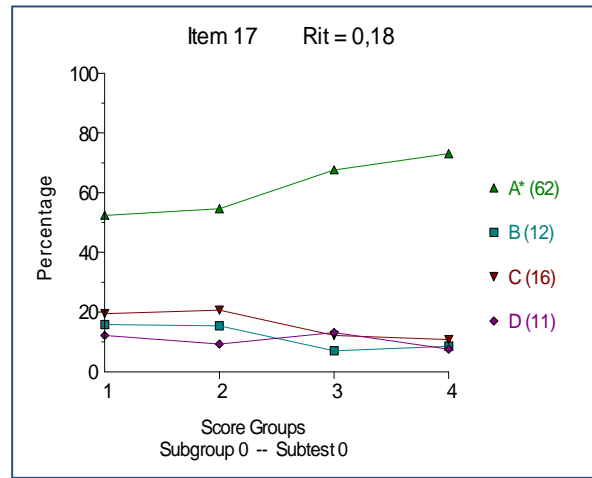
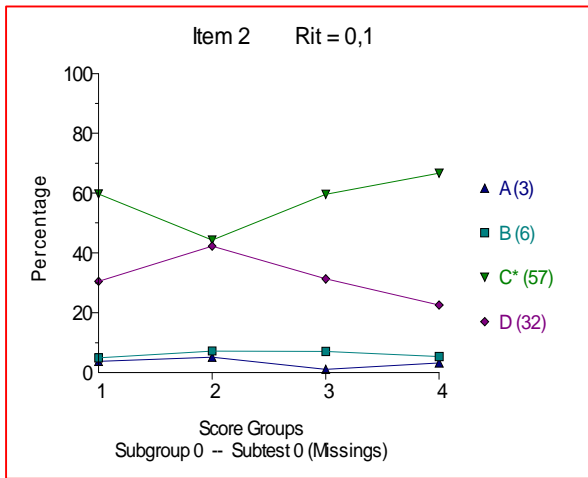


Table B.11B Poor or pathological items in PCP-Media Version B

item nr.	% of correct answer (p)	Rit	Rir	Flag code ¹	Graphical analysis
2	0,57	0,10	0,02	AB	This is poor because POOREST ones find the correct answer too easily. Guessing value HIGH
10	0,67	0,17	0,09	A	This is poor because POOREST ones find the correct answer too easily
20	0,61	0,18	0,10	A	The BEST ones do not find the correct answer; they are distracted by D. There seems to be two correct answers (A and D)
27	0,37	0,15	0,07	ABD	There is NO correct answer. The BEST ones are distracted by D. Check the key!
37	0,55	0,18	0,11	A	This is poor because POOREST ones find the correct answer too easily. Guessing value HIGH
47	0,71	0,13	0,06	A	The BEST ones do not find the correct answer they are distracted by D. There seems to be two correct answers (C and D)

- 1) A: Rit < 0.20 item-total correlation is low, B: Rar >= Rir a distracter correlates as high as or higher with the test's rest score than the correct alternative, C: Rir <= 0 the correct alternative does not correlate or even correlates negatively with the test's rest score, D: Rar >= 10 a distracter - test score correlation is suspiciously high



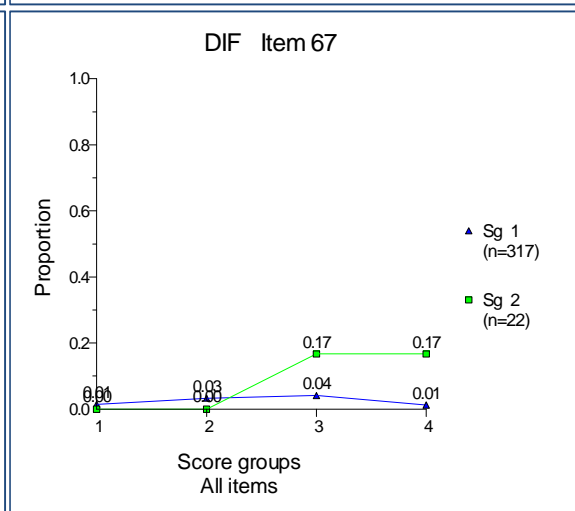
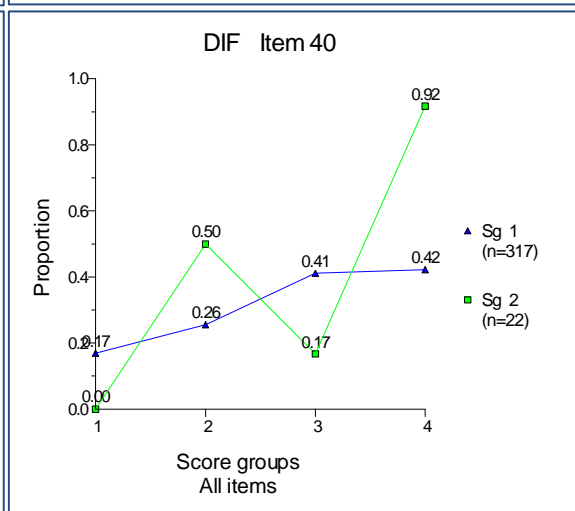
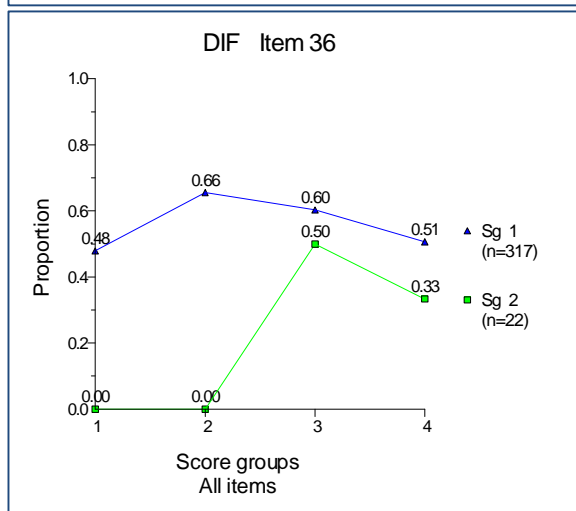
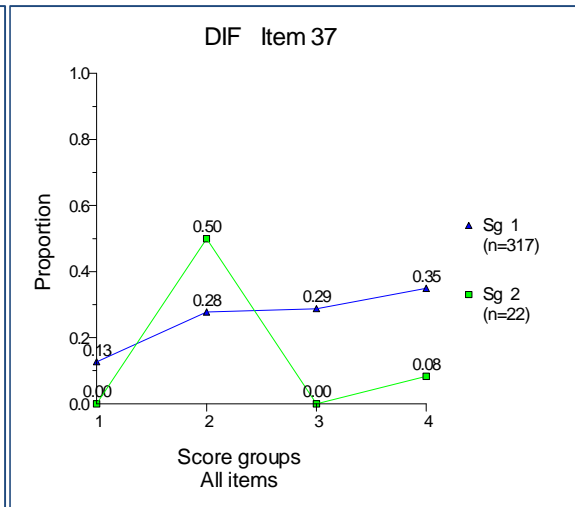
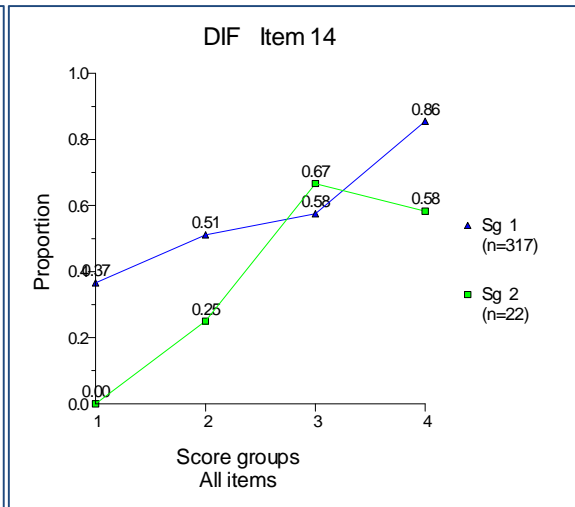
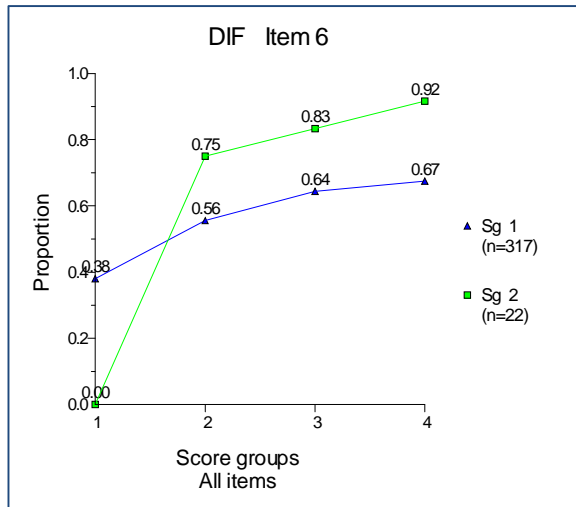
APPENDIX C

DIF analysis of the items in *INICÍA*

Table C.1A DIF of PCD-Básica Version A

Item ¹	DIF stat (MH)	z (standardized)	Item	DIF stat (MH)	z (standardized)	Item	DIF stat (MH)	z (standardized)	Item	DIF stat (MH)	z (standardized)
1	0	--	21	1,4575	0,1506	41	0,2811	-0,5302	61	1,0468	0,0311
2	1,2701	0,156	22	0,8913	-0,0632	42	1,3399	0,263	62	1,4541	0,34
3	0,2714	-0,4606	23	2,3077	0,3588	43	0,5269	-0,626	63	1,0169	0,0146
4	0,8411	-0,0953	24	--	--	44	1,2612	0,1708	64	1,9122	0,5551
5	0,5358	-0,3237	25	--	--	45	--	--	65	1,7843	0,3201
6	0,2838	-0,8539	26	2,2216	0,6353	46	0,7616	-0,1077	66	0,9804	-0,011
7	0,7753	-0,2498	27	--	--	47	1,5179	0,3723	67	0,1618	-1,0309
8	0,699	-0,3275	28	0,5343	-0,5587	48	0,4584	-0,5101	68	1,7162	0,2993
9	1,3032	0,1776	29	0,5954	-0,4932	49	0,8169	-0,1309	69	0,4101	-0,7751
10	0	--	30	0,3633	-0,6845	50	0,6183	-0,4561	70	1,0469	0,0428
11	0	--	31	--	--	51	1,0835	0,0517	71	0,6178	-0,4352
12	1,7532	0,4391	32	1,4634	0,3521	52	0,4347	-0,7574	72	0,7277	-0,2719
13	--	--	33	0,5282	-0,5839	53	--	--	73	0,6947	-0,3421
14	2,1452	0,7038	34	1,1548	0,1179	54	0,8464	-0,0916	74	0,4666	-0,6296
15	0,4358	-0,467	35	2,0605	0,3053	55	--	--	75	3,2743	0,8054
16	1,1432	0,1265	36	2,7297	0,9178	56	4,1384	0,5942	76	2,0529	0,647
17	--	--	37	2,8107	0,7141	57	1,3239	0,2526	77	0,967	-0,0303
18	0,7368	-0,2296	38	0,663	-0,2244	58	0,5511	-0,3997	78	--	--
19	--	--	39	--	--	59	0,5257	-0,5069	79	1,4182	0,3321
20	--	--	40	0,3662	-0,9285	60	1,2713	0,1328	80	1,7049	0,364

1) the items are highlighted when the $|z| > 0.700$



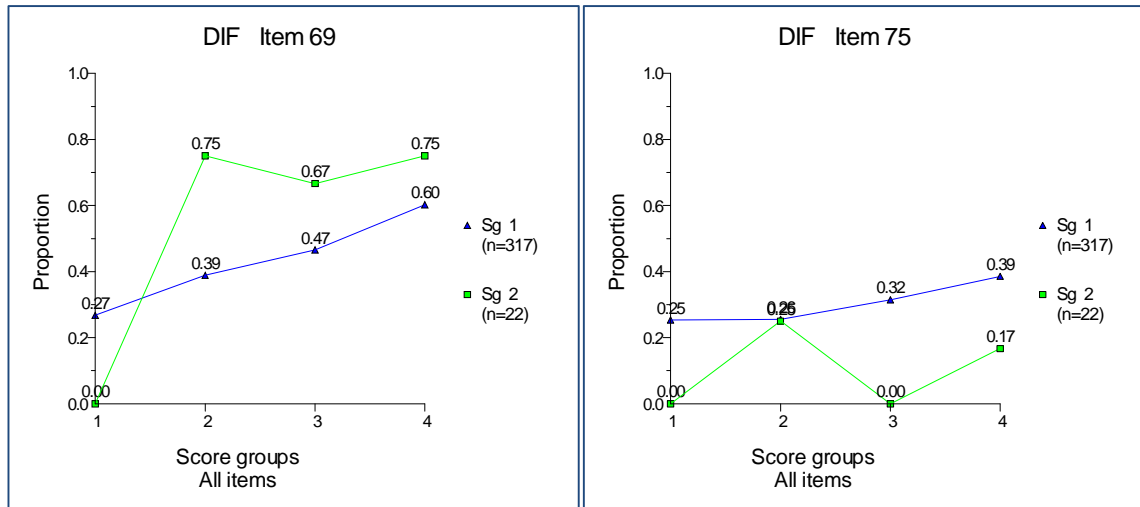


Table C.1B DIF of PCD-Básica Version B

Item ¹	DIF stat (MH)	z (standardized)	Item	DIF stat (MH)	z (standardized)	Item	DIF stat (MH)	z (standardized)	Item	DIF stat (MH)	z (standardized)
1	0,5223	-0,2608	21	0,2285	-0,6263	41	1,1186	0,0635	61	1,7846	0,4714
2	1,0992	0,0522	22	1,1446	0,0727	42	0,854	-0,1462	62	0,9332	-0,0622
3	0,5802	-0,358	23	0	--	43	1,0036	0,0032	63	0,8759	-0,1185
4	1,5483	0,2769	24	0,304	-0,4944	44	0	--	64	1,5112	0,3598
5	1,7243	0,4518	25	1,4174	0,285	45	2,0545	0,6395	65	1,0082	0,0053
6	1,3314	0,2499	26	0,9963	-0,0031	46	0,3074	-0,5996	66	0,8922	-0,0805
7	0,6955	-0,3263	27	4,0663	0,6041	47	1,1372	0,1164	67	1,3459	0,257
8	0,7725	-0,1908	28	0,1628	-1,0576	48	0,5764	-0,3746	68	2,0474	0,551
9	0,2182	-1,1259	29	0,7603	-0,2375	49	0,7946	-0,0976	69	1,1967	0,1654
10	0,4474	-0,3254	30	0,0923	-0,8155	50	0,5108	-0,5476	70	0,2428	-0,7299
11	1,1941	0,0959	31	1,0188	0,015	51	0,8019	-0,1934	71	0,8181	-0,1766
12	0,8576	-0,0846	32	0,9999	-0,0001	52	1,8074	0,5267	72	1,3145	0,2437
13	1,0326	0,0297	33	0,4738	-0,5984	53	0,9993	-0,0006	73	0,6966	-0,3269
14	0,9194	-0,0668	34	0,4374	-0,7491	54	2,3689	0,4281	74	0,8713	-0,117
15	0	--	35	1,358	0,1618	55	1,0642	0,0537	75	1,2917	0,2092
16	1,0963	0,0769	36	0,8246	-0,1499	56	1,6214	0,414	76	0,738	-0,2805
17	0,8533	-0,1146	37	0,1754	-0,8536	57	0,9059	-0,0887	77	1,6193	0,4459
18	0,9513	-0,0363	38	0,473	-0,5863	58	2,747	0,6422	78	1,3606	0,2768
19	1,3283	0,2394	39	0,7337	-0,2789	59	0,9627	-0,0322	79	0,9923	-0,0065
20	0,8749	-0,1214	40	0,4601	-0,711	60	0,9645	-0,0291	80	0,4753	-0,3006

1) the items are highlighted when the $|z| > 0.700$

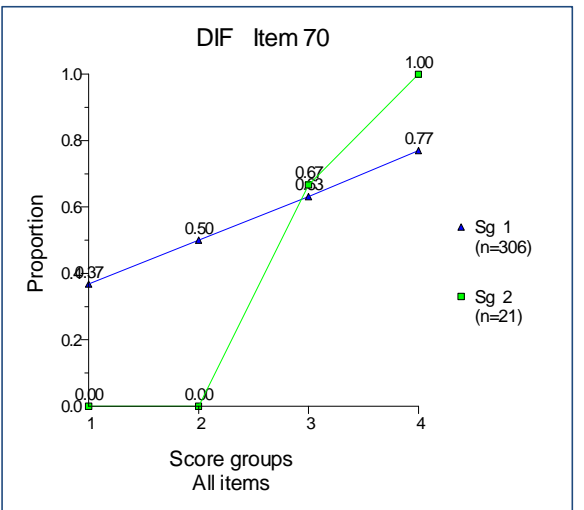
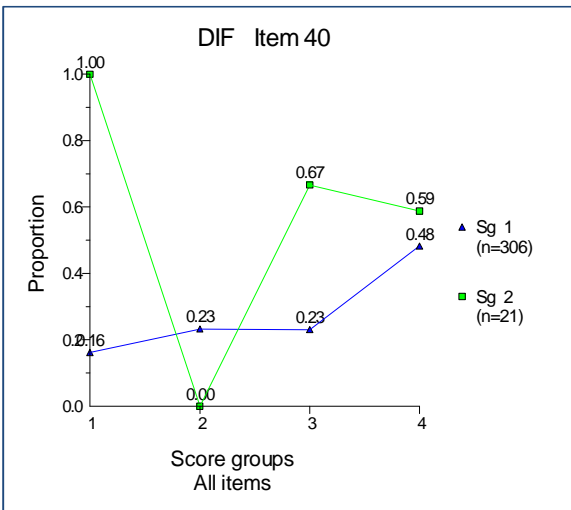
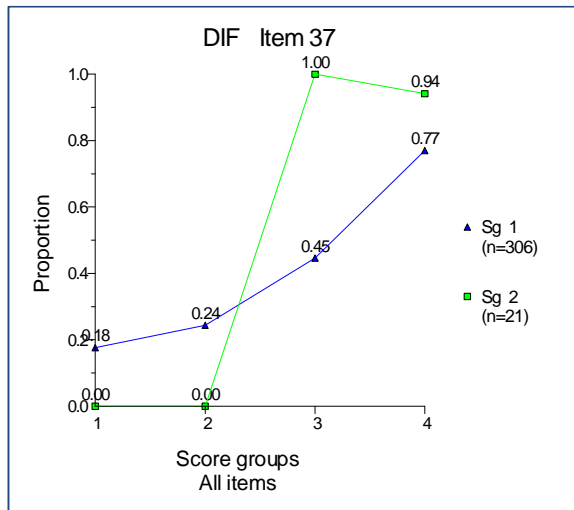
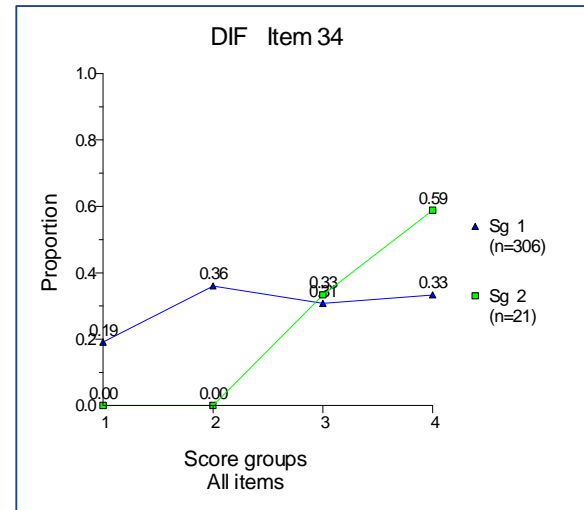
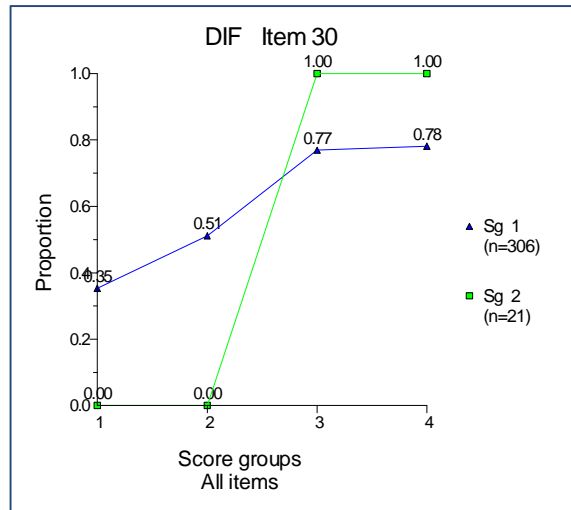
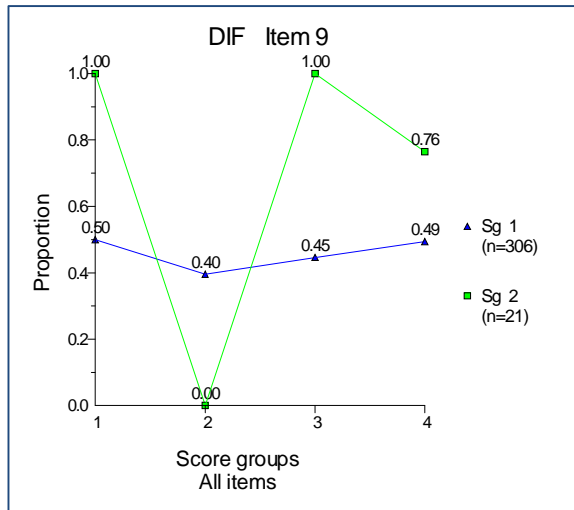
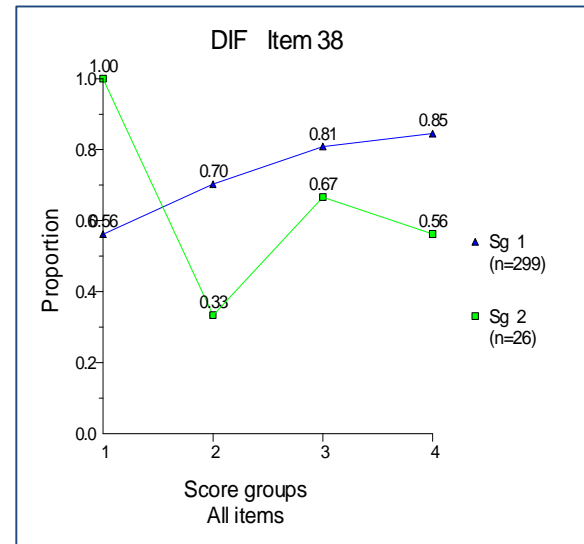
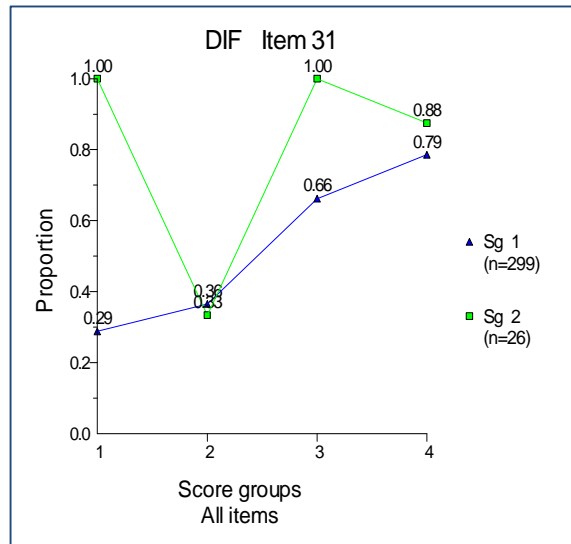
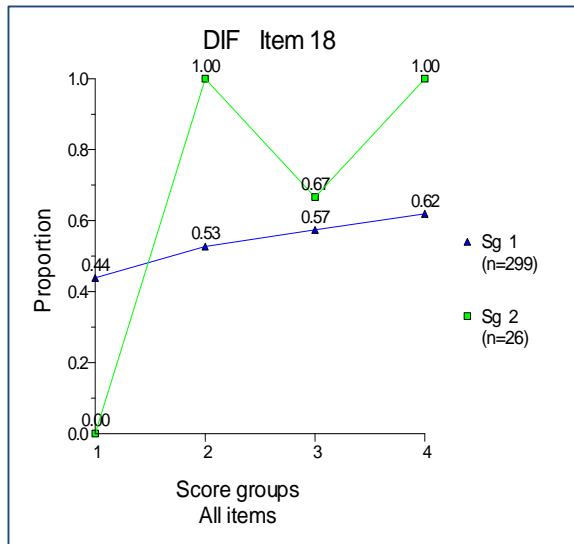
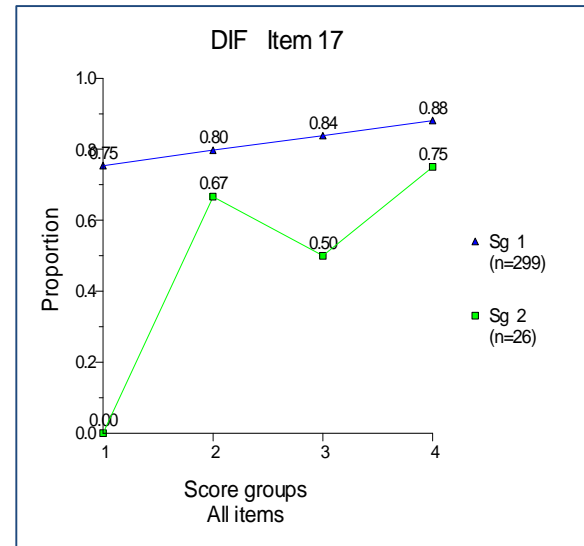
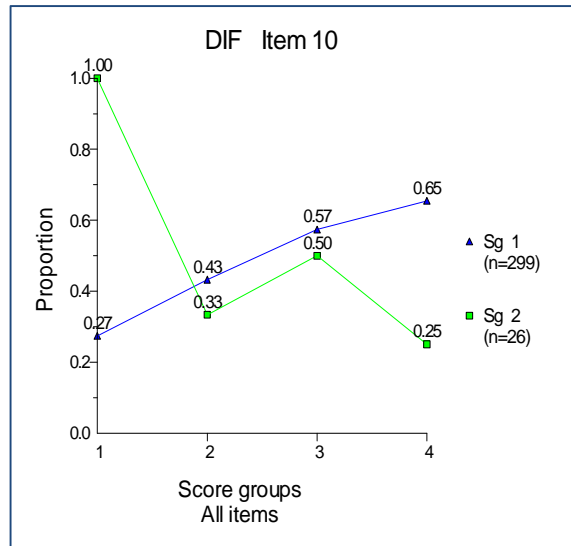
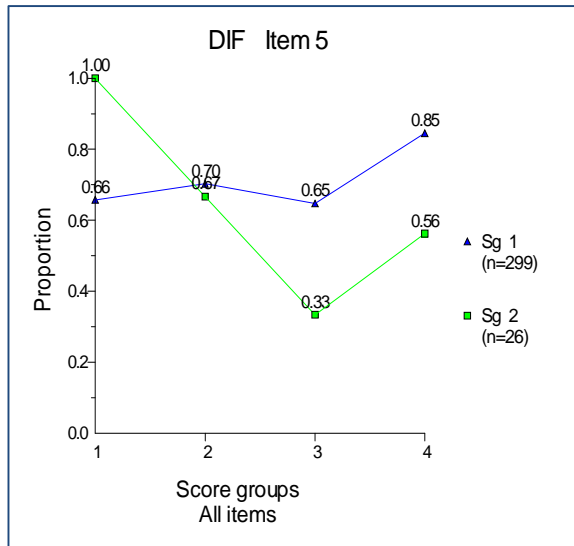


Table C.2A DIF of PCP Básica Version A

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	0,6878	-0,3646	21	1,2879	0,1028	41	0,9494	-0,0303
2	0,9458	-0,0565	22	1,1623	0,1486	42	0,1122	-0,9263
3	1,6965	0,3481	23	1,6181	0,4219	43	1,0397	0,0316
4	0,7925	-0,2011	24	0	--	44	0,6616	-0,361
5	2,9413	1,0705	25	0	--	45	1,077	0,067
6	1,5691	0,3828	26	0,633	-0,3353	46	0,7972	-0,2158
7	0,5798	-0,5474	27	0,634	-0,3005	47	1,9573	0,627
8	1,4015	0,3354	28	0,8317	-0,1632	48	2,68	0,748
9	0,4091	-0,375	29	1,072	0,044	49	1,0078	0,008
10	2,5052	0,9358	30	0,5296	-0,5652	50	0,4494	-0,7321
11	2,1165	0,6939	31	0,3547	-0,7512			
12	0,8231	-0,1497	32	1,0914	0,0766			
13	0,7508	-0,2854	33	0,7428	-0,2153			
14	0,6641	-0,3291	34	1,9044	0,5816			
15	0,4299	-0,5781	35	0,9938	-0,0035			
16	0,6831	-0,3149	36	0,9125	-0,0908			
17	3,3309	1,1125	37	1,3219	0,2645			
18	0,1846	-1,144	38	3,0133	1,09			
19	0	--	39	0,909	-0,1019			
20	1,1439	0,1366	40	1,1496	0,1381			

1) the items are highlighted when the $|z| > 0.700$



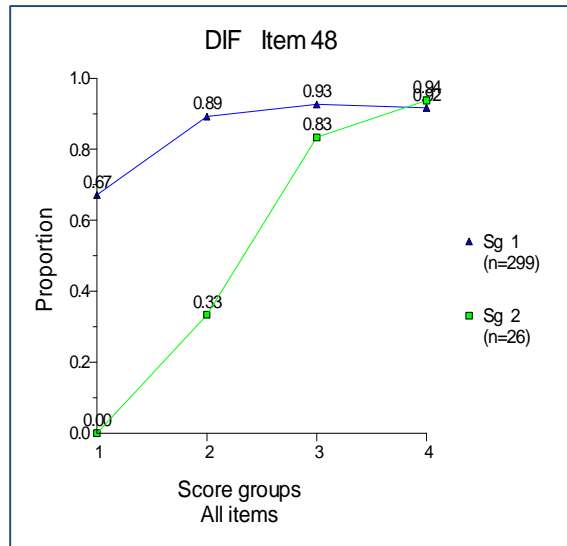
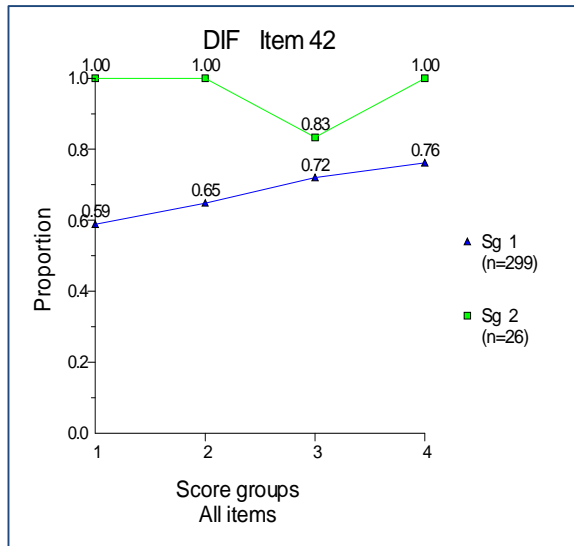


Table C.2B DIF of PCP-Básica Version B

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	1,6074	0,401	21	1,2907	0,2102	41	0,7086	-0,263
2	0,5989	-0,4266	22	0,2753	-0,72	42	0,3317	-0,468
3	0,5671	-0,4836	23	1,6906	0,4662	43	0,352	-0,5596
4	0,2511	-0,7614	24	0,502	-0,4083	44	0,7596	-0,254
5	0,6305	-0,3086	25	4,3894	0,9012	45	1,6566	0,3139
6	0,5354	-0,3283	26	0,9099	-0,0655	46	0,9541	-0,0301
7	2,3509	0,7306	27	1,0039	0,0031	47	0,4772	-0,4914
8	0,6222	-0,4211	28	1,0359	0,025	48	1,7634	0,3646
9	0,6583	-0,2352	29	1,0204	0,0128	49	0,6225	-0,4002
10	0,9147	-0,0735	30	0,2123	-0,8942	50	2,0499	0,5152
11	1,3509	0,2329	31	0,7322	-0,2083			
12	0,43	-0,5166	32	0,6242	-0,1942			
13	1,0567	0,0429	33	1,5671	0,3979			
14	0,2317	-0,5827	34	0,1862	-1,3682			
15	1,3587	0,2355	35	0,98	-0,0153			
16	0,5354	-0,4188	36	1,627	0,3687			
17	0,9796	-0,0157	37	0,674	-0,2707			
18	1,5583	0,3834	38	0,2608	-0,5286			
19	2,3691	0,4569	39	1,3232	0,2294			
20	1,8621	0,4744	40	0,7015	-0,3046			

1) the items are highlighted when the $|z| > 0.700$

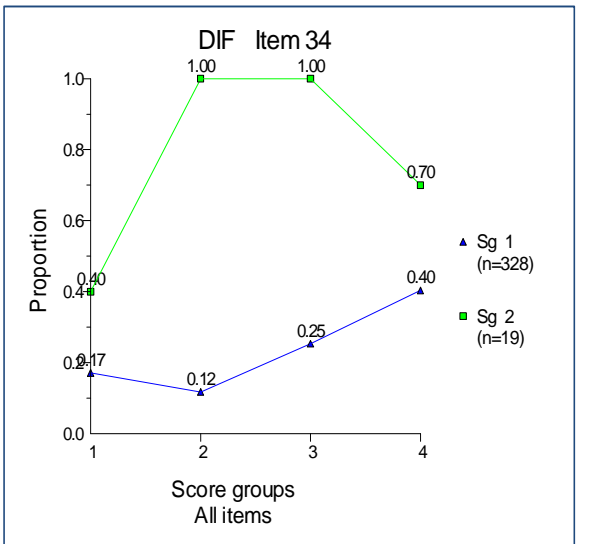
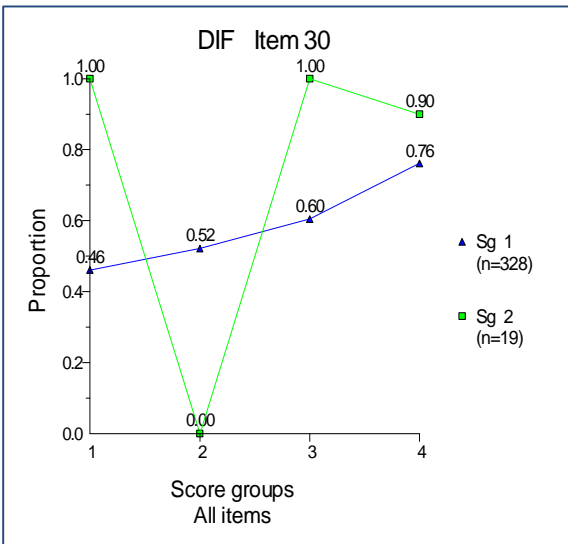
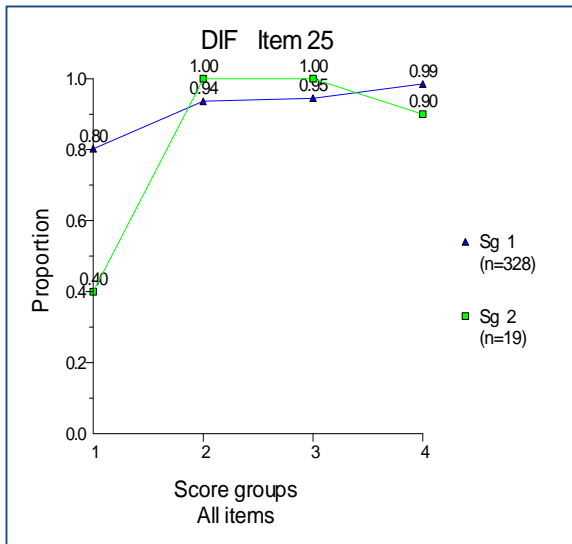
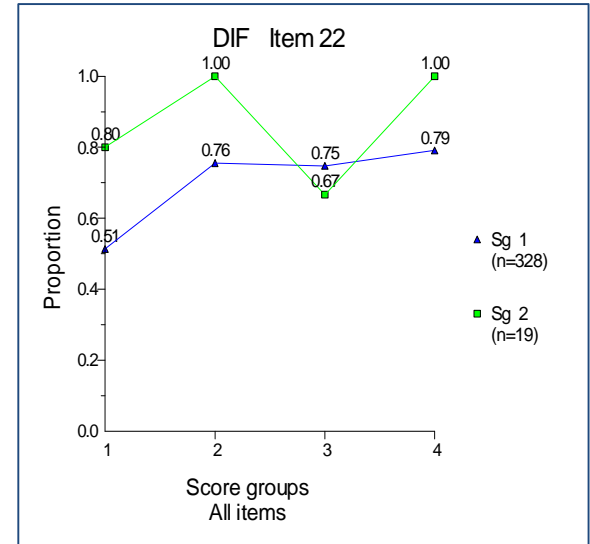
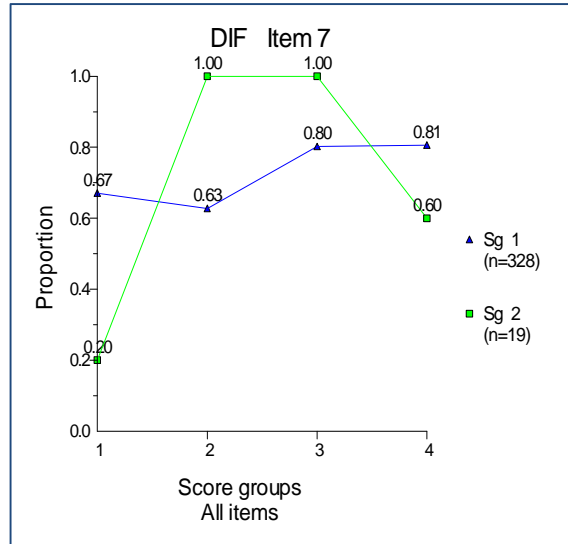
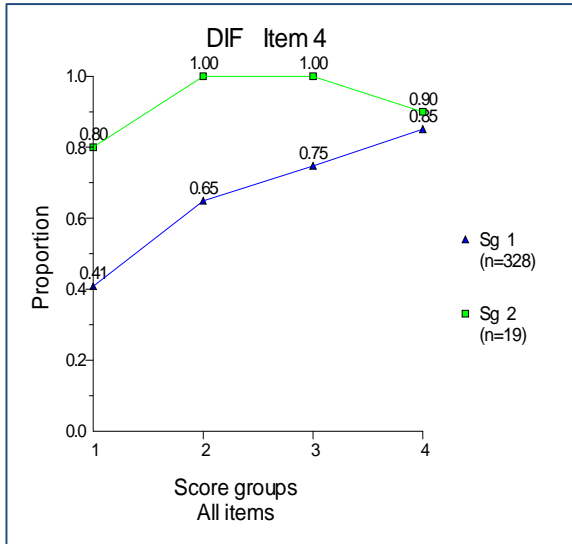
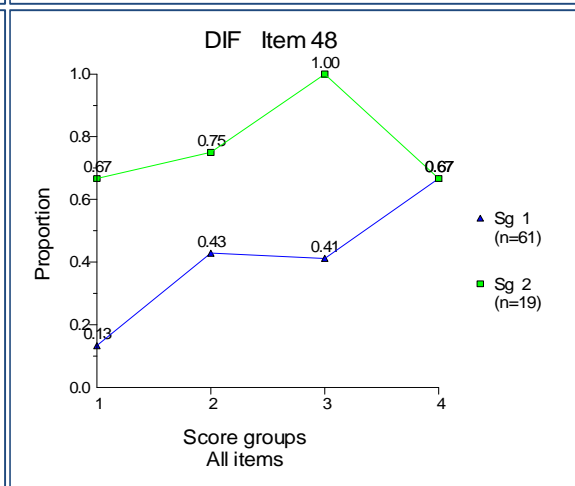
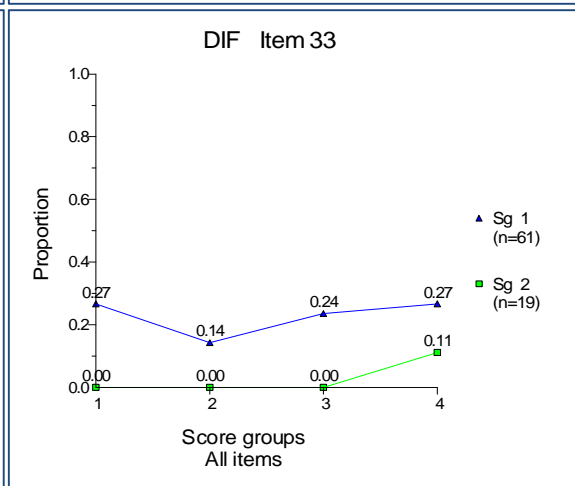
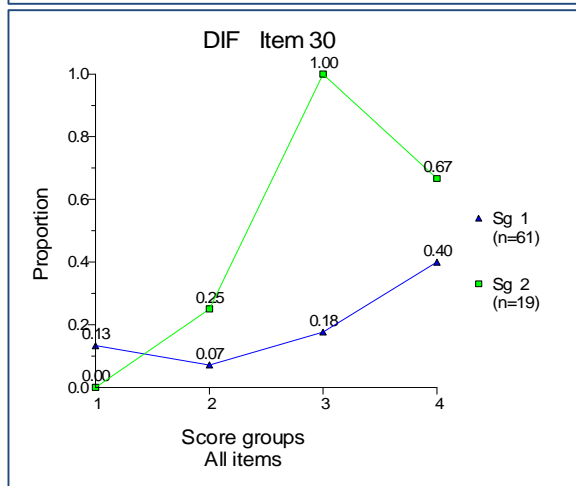
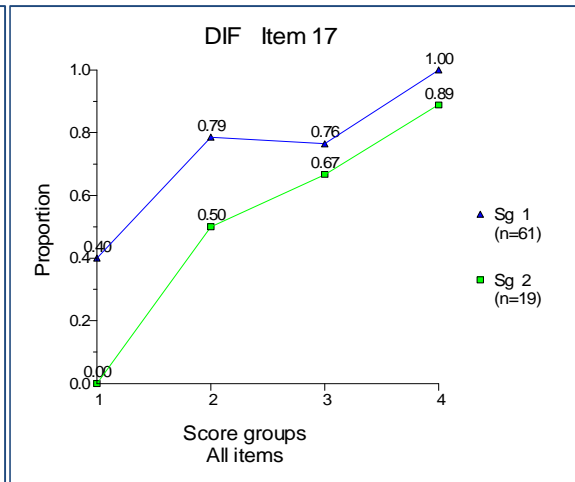
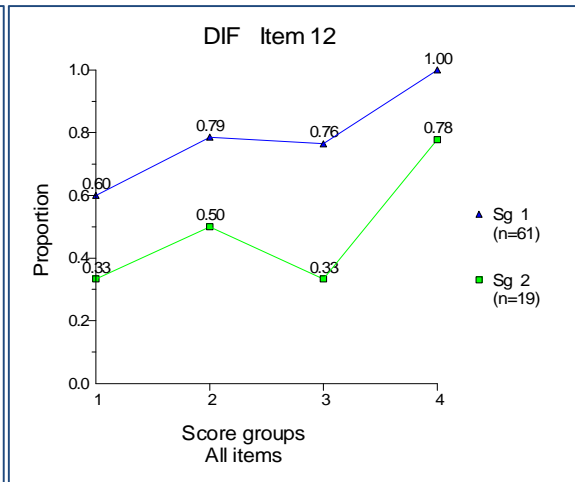
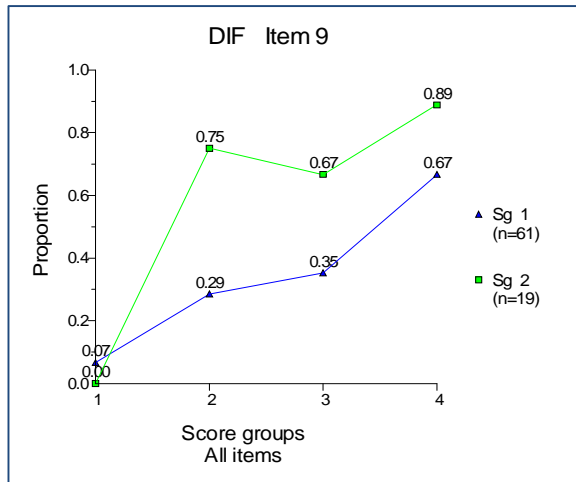


Table C.3 DIF of *PCD-Biología*

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	1,764	0,3392	21	0,8241	-0,1401	41	0,685	-0,2768
2	1,5544	0,3473	22	1,5153	0,3028	42	1,4564	0,2613
3	1,5865	0,3522	23	0,7981	-0,1602	43	0,4611	-0,2955
4	0,839	-0,1135	24	0	--	44	0,5292	-0,3335
5	1,4456	0,2258	25	1,4066	0,1852	45	1,0294	0,0191
6	0,8899	-0,0829	26	0,9611	-0,0281	46	0,755	-0,2051
7	3,3231	0,6887	27	0,4444	-0,46	47	--	--
8	2,4856	0,613	28	1,4962	0,2391	48	0,3065	-0,8641
9	0,2494	-0,8796	29	1,1031	0,0707	49	0,4809	-0,4051
10	0,8112	-0,1648	30	0,2464	-0,9938	50	0,7577	-0,2173
11	0,3371	-0,4346	31	0,7341	-0,2487	51	0,3223	-0,8694
12	5,5064	1,0497	32	1,455	0,261	52	0,5447	-0,4958
13	0,8489	-0,1318	33	6,6424	0,7137	53	0,8053	-0,1244
14	0,3172	-0,6207	34	0,8037	-0,15	54	0,7483	-0,1026
15	0,6823	-0,2513	35	0	--	55	0,6263	-0,3514
16	1,0244	0,0191	36	1,3583	0,1745	56	1,3898	0,224
17	4,7689	0,8523	37	1,0388	0,0228	57	0,5455	-0,4778
18	1,8195	0,3922	38	1,3192	0,1907	58	0,9723	-0,022
19	0,5367	-0,4422	39	0,6264	-0,3404	59	1,289	0,1909
20	1,1309	0,0819	40	0,625	-0,1673	60	0,5498	-0,4457

1) the items are highlighted when the $|z| > 0.700$



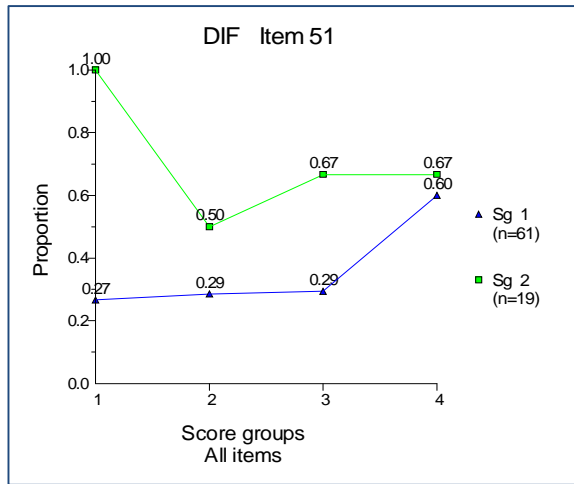
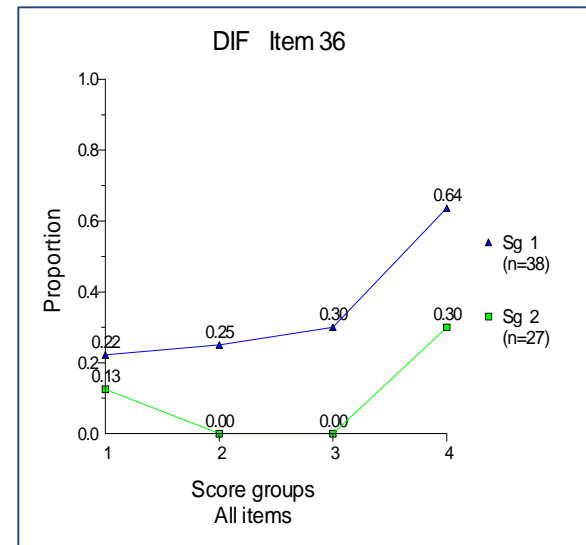
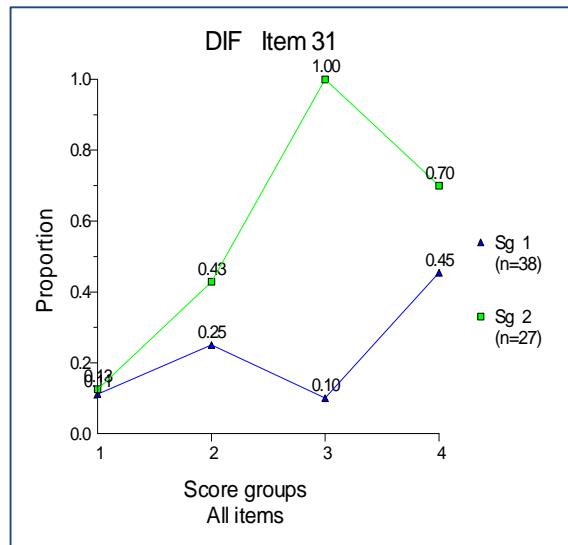
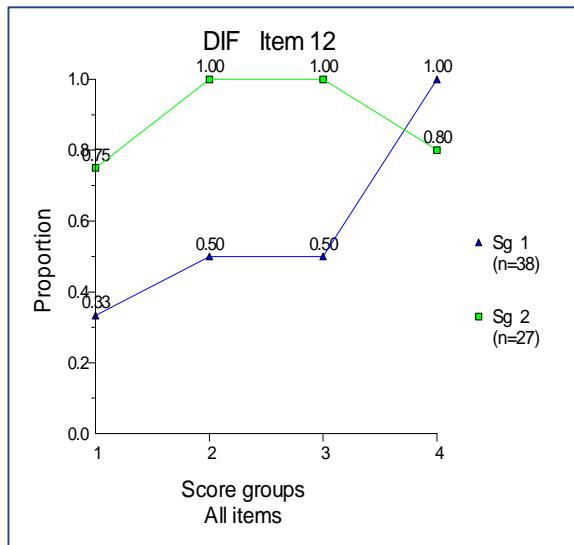
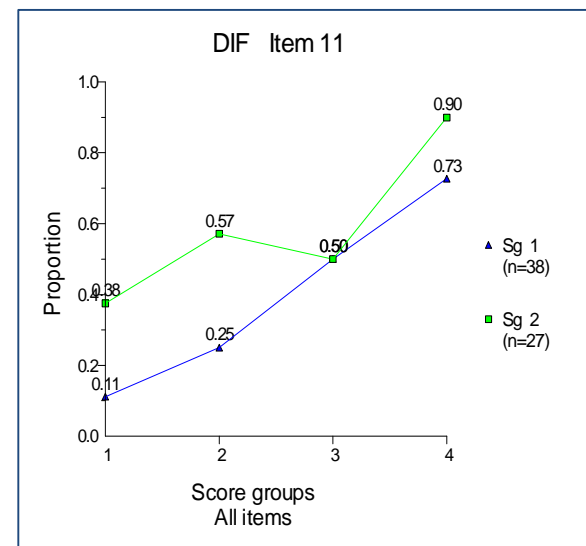
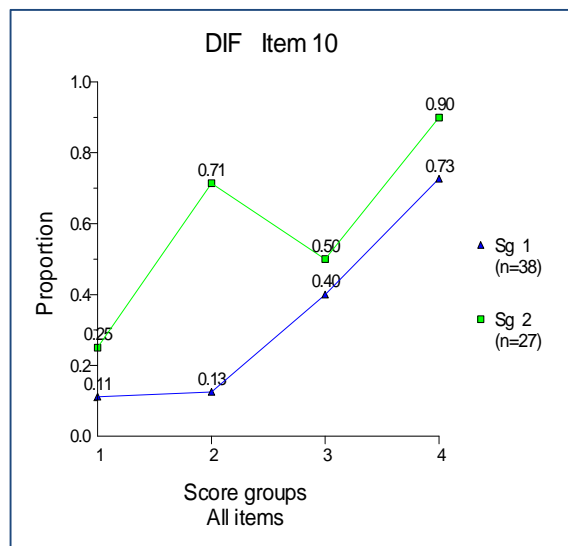
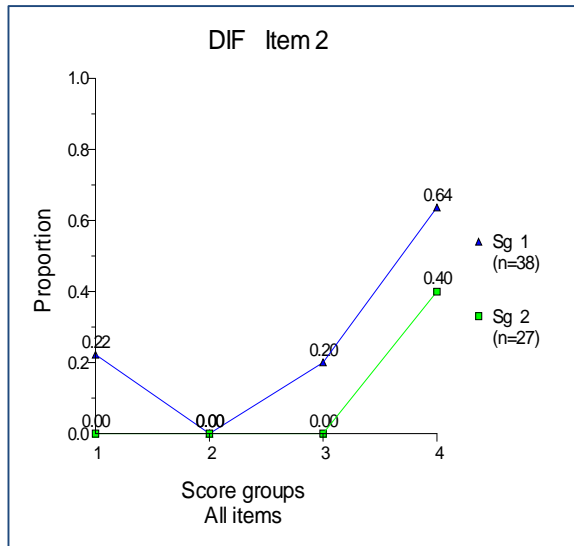


Table C.4 DIF of PCD-Física

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	1,4289	0,2733	21	0,9134	-0,0587	41	0,5624	-0,4181
2	4,2978	0,7525	22	0,6488	-0,2359	42	2,4597	0,6283
3	1,3561	0,2098	23	0,7654	-0,1718	43	1,5946	0,2472
4	1,4228	0,2322	24	1,9812	0,4117	44	0,804	-0,1823
5	1,0111	0,0082	25	1,6504	0,3898	45	0,4582	-0,6249
6	1,584	0,2895	26	1,0496	0,0381	46	1,1249	0,1001
7	3,5347	0,6266	27	1,2154	0,145	47	1,532	0,2587
8	0,7538	-0,2174	28	0,9741	-0,0197	48	1,5285	0,2835
9	0,8585	-0,103	29	0,7813	-0,1664	49	0,8349	-0,13
10	0,2373	-0,9591	30	0,8991	-0,083	50	0,3314	-0,7759
11	0,3164	-0,7975	31	0,3209	-0,8465	51	1,4608	0,3015
12	0,2907	-0,8272	32	0,4624	-0,4988	52	0,9906	-0,0067
13	1,8485	0,4501	33	1,0447	0,0284	53	2,1588	0,6117
14	1,3582	0,2398	34	0,5787	-0,3963	54	2,6412	0,7526
15	0,6192	-0,3278	35	0,9588	-0,0343	55	0,8172	-0,1569
16	1,1456	0,0942	36	4,6687	0,9163	56	1,1453	0,0776
17	1,5915	0,3492	37	2,218	0,5819	57	0,5727	-0,4333
18	0,5572	-0,4427	38	4,3463	1,0789	58	1,029	0,0208
19	2,4514	0,6784	39	2,208	0,6238	59	1,396	0,2594
20	1,0943	0,0605	40	1,12	0,0725	60	0,2603	-0,8879

1) the items are highlighted when the $|z| > 0.700$



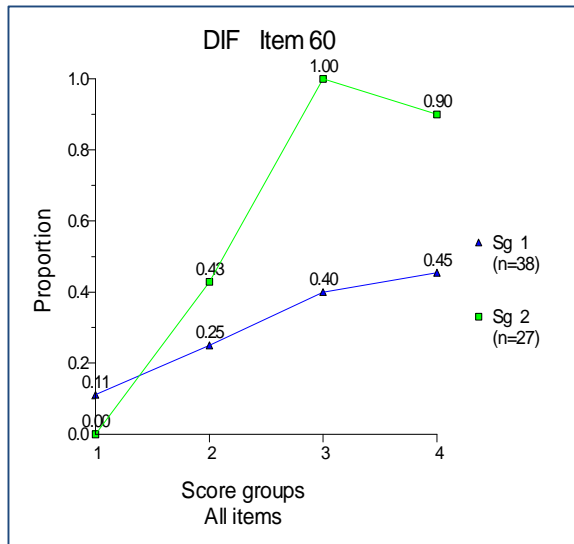
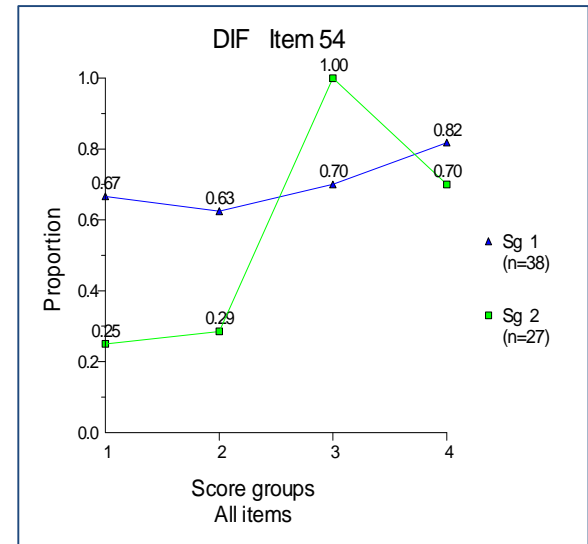
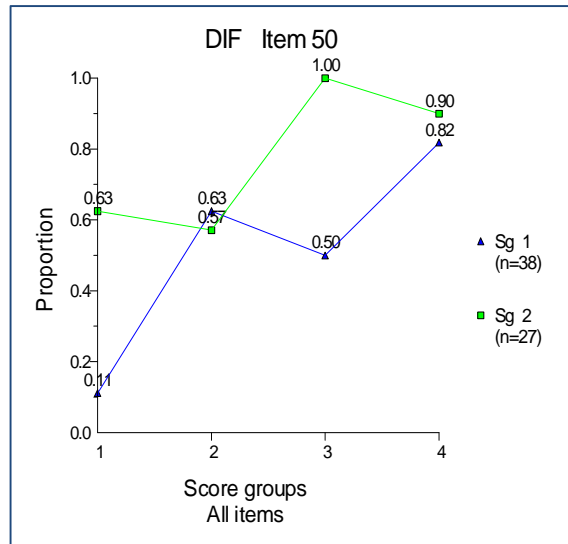
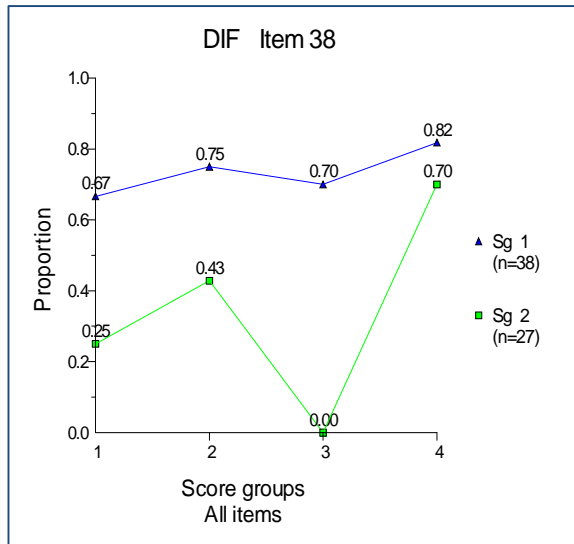


Table C.5 DIF of PCD-Matemática

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	1,168	0,0933	21	1,1803	0,196	41	1,1501	0,1819
2	0,3298	-0,6507	22	1,6261	0,5967	42	1,3646	0,3431
3	1,1004	0,0691	23	1,4851	0,419	43	1,3001	0,3137
4	0,8857	-0,1362	24	0,6772	-0,4806	44	0,9156	-0,118
5	0,9132	-0,0957	25	0,6929	-0,5012	45	1,1387	0,1539
6	0,7823	-0,3168	26	1,4671	0,4209	46	0,4894	-0,8814
7	0,7899	-0,3009	27	0,9112	-0,096	47	0,8433	-0,2287
8	1,8241	0,814	28	2,2562	0,9856	48	1,067	0,0854
9	1,314	0,3707	29	0,5673	-0,635	49	0,5846	-0,5258
10	1,0745	0,1007	30	0,2828	-1,0924	50	0,7058	-0,4314
11	1,4829	0,5073	31	0,7347	-0,3973	51	1,2688	0,324
12	0,9667	-0,042	32	0,9909	-0,0075	52	0,8034	-0,2863
13	0,8043	-0,2443	33	0,7507	-0,3848	53	0,8355	-0,2404
14	0,8467	-0,1988	34	0,9914	-0,0109	54	1,3551	0,3981
15	0,9593	-0,0452	35	1,049	0,0637	55	1,2111	0,2404
16	1,0426	0,0536	36	1,1281	0,1377	56	0,5897	-0,6884
17	1,0672	0,0786	37	1,2634	0,339	57	1,3751	0,3585
18	1,4997	0,5362	38	0,3865	-1,171	58	0,541	-0,7691
19	0,9474	-0,064	39	0,5812	-0,6956	59	0,6625	-0,565
20	0,7698	-0,3447	40	1,5323	0,5157	60	1,0749	0,0966

1) the items are highlighted when the $|z| > 0.700$

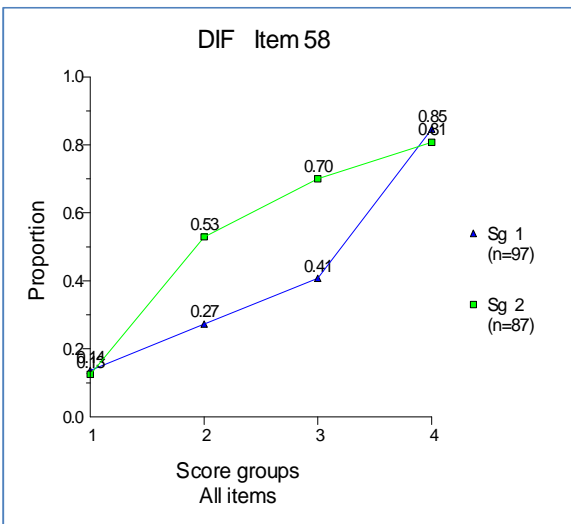
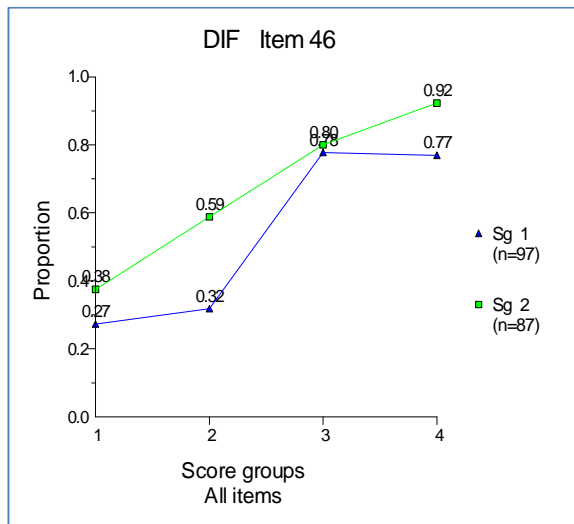
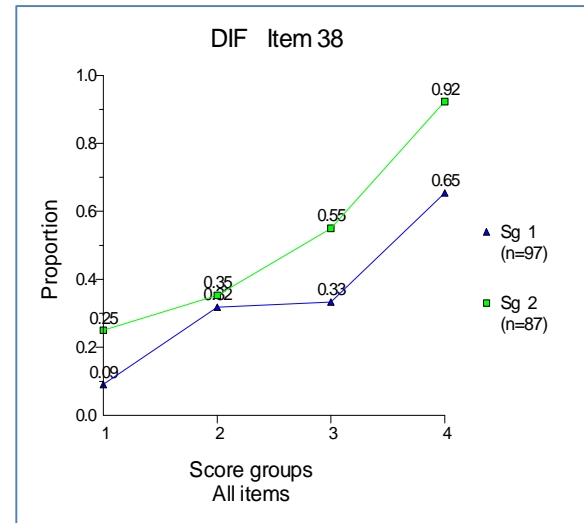
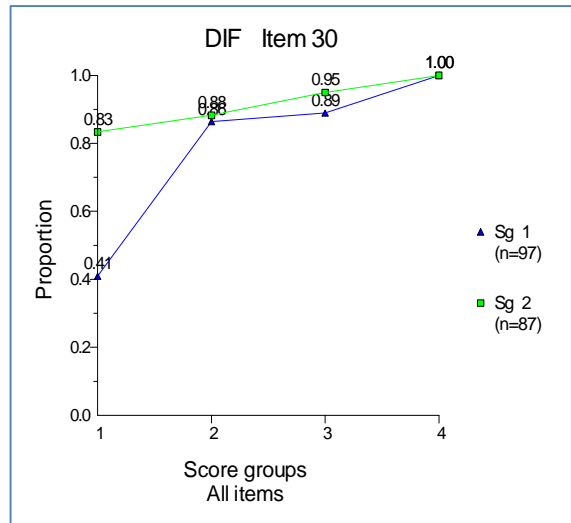
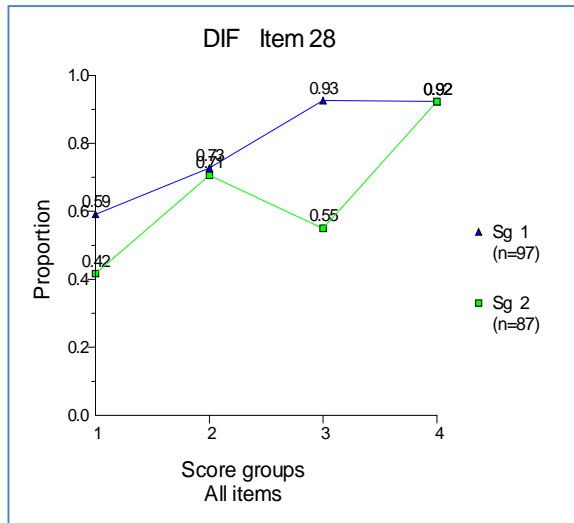


Table C.6 DIF of PCD-Química

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	1,9636	0,2494	21	0,3649	-0,4035	41	0,846	-0,0875
2	0,7778	-0,0988	22	2,1529	0,3561	42	1,2427	0,115
3	2,3867	0,4196	23	0,8482	-0,0587	43	0,1751	-0,9147
4	0,3857	-0,3914	24	0,4023	-0,5087	44	2,3978	0,486
5	1,7104	0,2939	25	5,4468	0,8018	45	--	--
6	0,5455	-0,2958	26	2,1481	0,3373	46	3,4648	0,6002
7	0,7453	-0,1617	27	1,4519	0,1856	47	1,4786	0,2022
8	1,7284	0,2064	28	1,8511	0,3276	48	0,5414	-0,3197
9	0,9487	-0,0314	29	3,7884	0,6731	49	0,2165	-0,6475
10	0,6892	-0,1964	30	1,5799	0,2388	50	2,6056	0,4431
11	0,5769	-0,2392	31	1,8611	0,2578	51	1,5813	0,2636
12	0,3199	-0,5416	32	1,2014	0,0626	52	0,1964	-0,8077
13	0,4367	-0,4023	33	0,7563	-0,158	53	1,3333	0,097
14	0,2705	-0,6329	34	3,2014	0,4523	54	0,5586	-0,2185
15	1,0056	0,0029	35	0	--	55	0,3889	-0,3518
16	1,3116	0,1596	36	2,4848	0,4583	56	1,6071	0,2767
17	1,3333	0,1613	37	0,3529	-0,3274	57	0,5652	-0,3182
18	0,4583	-0,4106	38	0,6207	-0,2257	58	0,5389	-0,3304
19	1,0569	0,0294	39	0	--	59	0,3862	-0,4924
20	0,9407	-0,0271	40	1,1217	0,0694	60	0,8596	-0,0752

1) the items are highlighted when the $|z| > 0.700$

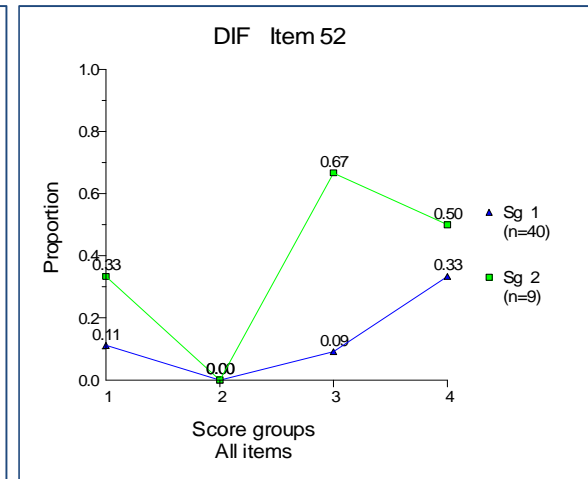
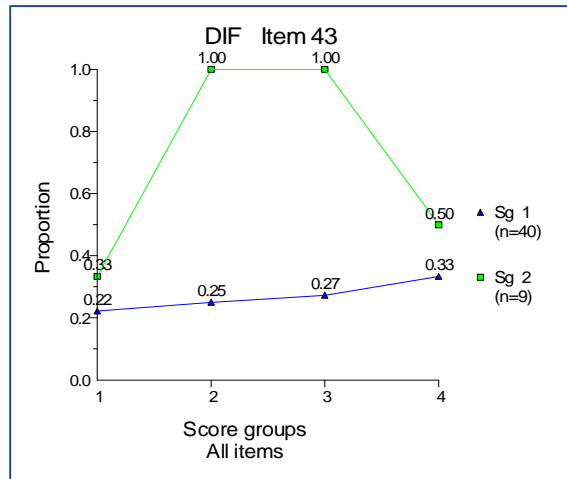
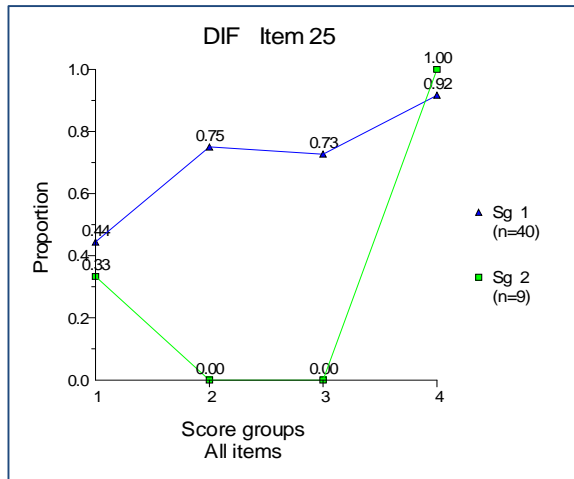


Table C.7A DIF of PCD-Historia Version A

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	31,8064	1,1923	21	1,6217	0,2756	41	0	--
2	0,6699	-0,2157	22	1,1101	0,0593	42	1,7115	0,21
3	0,8772	-0,0634	23	1,281	0,1464	43	2,1111	0,3947
4	1,7827	0,3012	24	0,3131	-0,4424	44	0	--
5	0	--	25	2,9947	0,5238	45	0,6659	-0,1976
6	1,1198	0,065	26	2,175	0,4133	46	0,6482	-0,2153
7	0,7231	-0,1881	27	1,3742	0,1534	47	1,2596	0,1067
8	2,0936	0,3725	28	0	--	48	1,2723	0,1305
9	1,237	0,1112	29	1,7054	0,2494	49	1,9226	0,361
10	0,8444	-0,0939	30	0	--	50	1,3425	0,1639
11	2,726	0,5932	31	0,24	-0,6727	51	0,8019	-0,1258
12	4,2196	0,7845	32	0,9377	-0,036	52	1,9096	0,3376
13	0	--	33	0,3592	-0,481	53	7,5302	1,0131
14	0,9403	-0,0261	34	0	--	54	1,7724	0,3052
15	1,7947	0,3292	35	0	--	55	0,444	-0,3856
16	1,2186	0,1146	36	0	--	56	0,7698	-0,0988
17	1,5736	0,2675	37	0,1831	-0,5722	57	1,7601	0,3003
18	0,8455	-0,0636	38	0,5733	-0,2157	58	0	--
19	3,0795	0,4508	39	1,9153	0,3741	59	0,8973	-0,0574
20	0	--	40	0,893	-0,0468	60	--	--

1) the items are highlighted when the $|z| > 0.700$

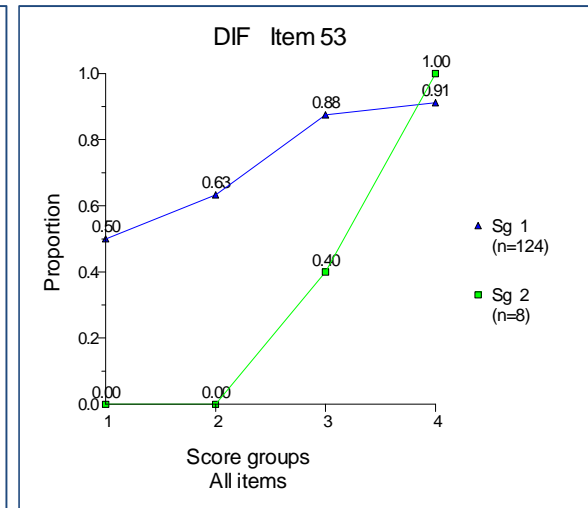
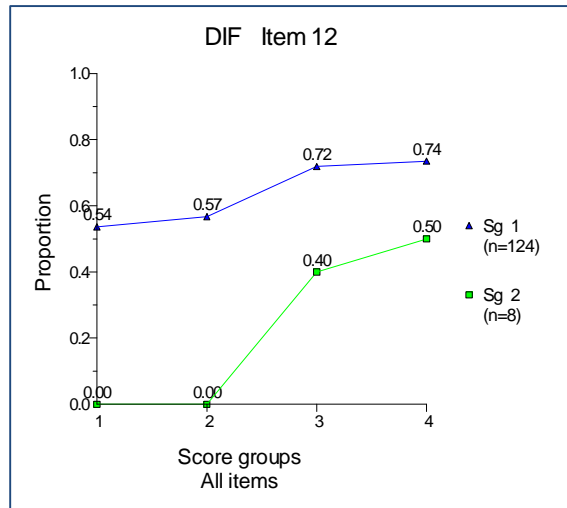
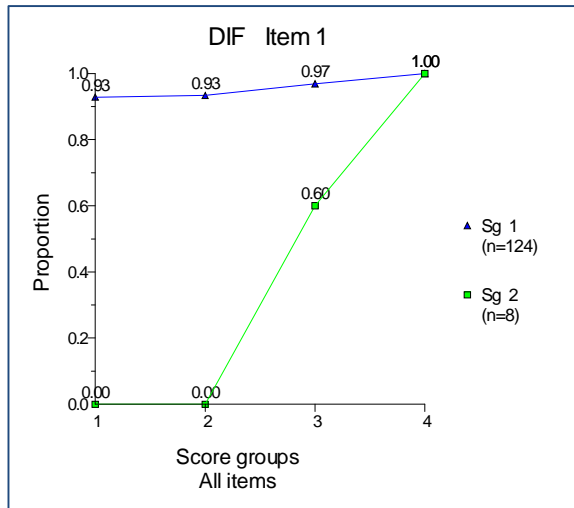


Table C.7B DIF of PCD-Historia Version B

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	0,5072	-0,2714	21	35,4359	1,0558	41	0,8903	-0,0444
2	1,8667	0,2928	22	0,397	-0,334	42	0	--
3	0	--	23	0,858	-0,0718	43	1,8508	0,2028
4	1,1432	0,0626	24	0	--	44	3,5357	0,6088
5	4,6681	0,808	25	0,4245	-0,3383	45	1,8007	0,2768
6	0	--	26	2,6292	0,432	46	1,2814	0,1357
7	0	--	27	2,1076	0,3449	47	0	--
8	2,4545	0,2673	28	6,5178	0,7686	48	0	--
9	3,8766	0,5939	29	0,4716	-0,2978	49	1,0491	0,0185
10	1,109	0,0552	30	0,8451	-0,0886	50	0,8196	-0,107
11	0	--	31	0	--	51	2,0578	0,3895
12	0	--	32	3,7758	0,7068	52	1,0471	0,0214
13	4,8444	0,7059	33	0,4436	-0,4128	53	0	--
14	1,0989	0,0446	34	1,1434	0,0722	54	2,0253	0,2581
15	5,8333	0,6869	35	1,0741	0,0374	55	0	--
16	1,1434	0,0722	36	1,202	0,0952	56	0,8568	-0,0563
17	21,4125	0,9257	37	0,3673	-0,3887	57	0,5272	-0,2726
18	2,5929	0,5125	38	0	--	58	1,0941	0,0419
19	1,0737	0,0379	39	0	--	59	0	--
20	8,5	0,6156	40	0	--	60	0,3875	-0,3805

1) the items are highlighted when the $|z| > 0.700$

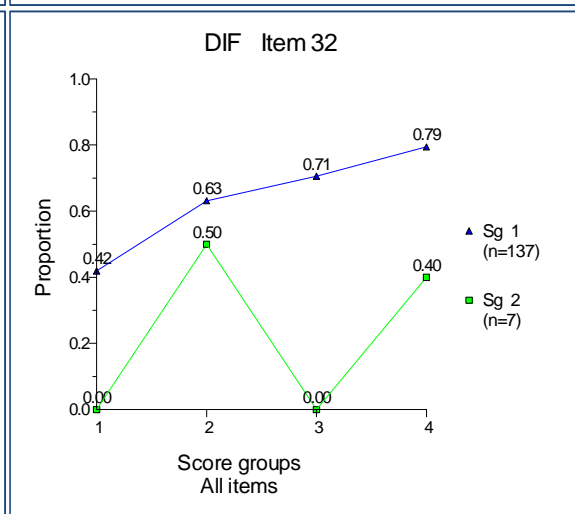
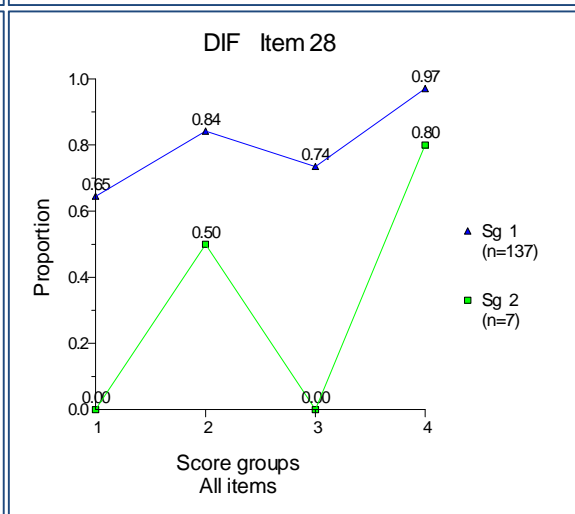
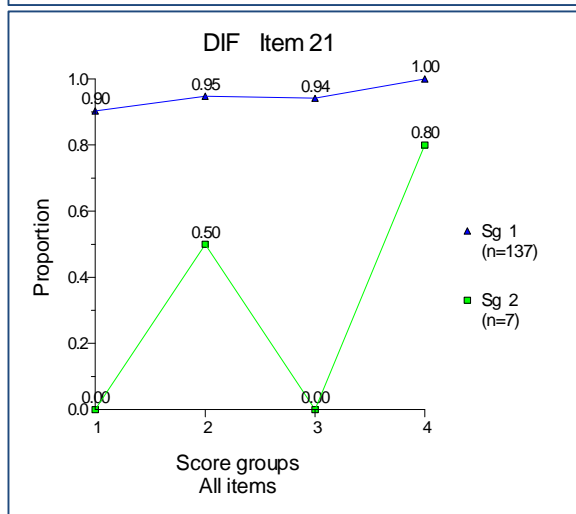
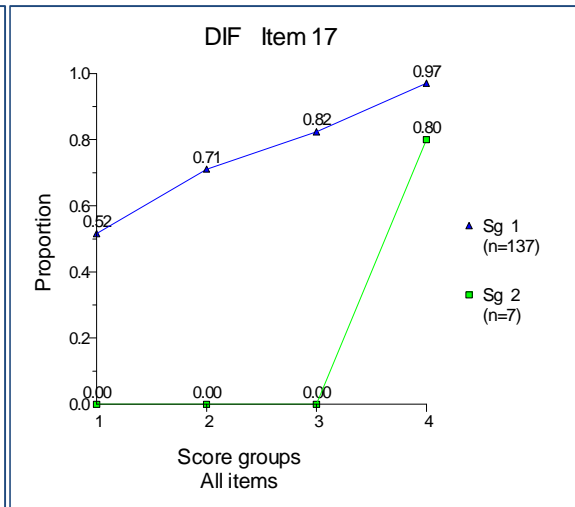
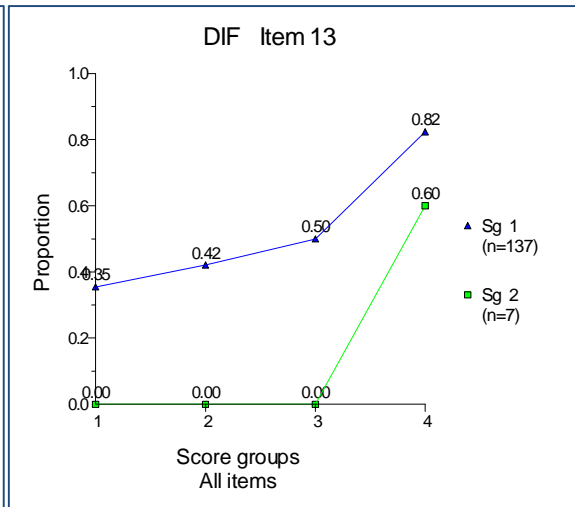
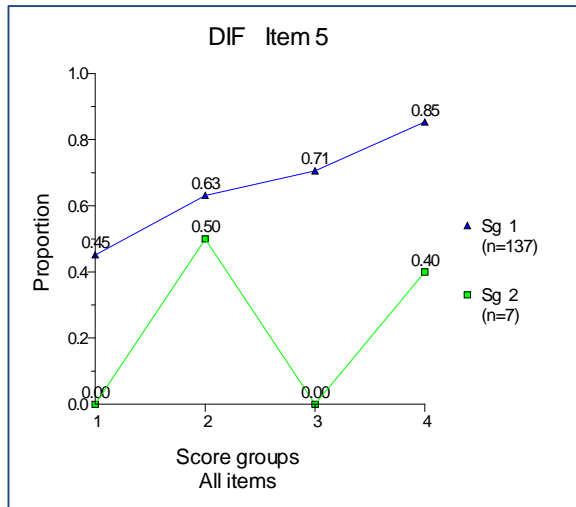


Table C.8A DIF of PCD-Lenguaje Version A

Item ¹	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z
1	0	--		21	0,7373		-0,2098	41	1,1447		0,0732
2	2,1482	0,4943		22	0,2661		-0,6815	42	1,1372		0,0866
3	2,2669	0,4439		23	0,7385		-0,206	43	1,4462		0,2465
4	0,369	-0,573		24	5,6667		0,7082	44	0,2086		-0,7877
5	2,1579	0,5401		25	0,6698		-0,2435	45	1,3972		0,1604
6	1,0974	0,0595		26	1,3776		0,2169	46	0		--
7	0,7947	-0,1485		27	2,6539		0,6125	47	0,9511		-0,027
8	0,7376	-0,2009		28	2,7157		0,6194	48	0,9502		-0,0284
9	1,0832	0,0512		29	2,2523		0,5422	49	1,2953		0,1692
10	1,5128	0,2695		30	0,7306		-0,2112	50	0,3966		-0,4502
11	2,2379	0,5021		31	1,8208		0,3679	51	2,4862		0,6221
12	1,9775	0,4389		32	1,7692		0,391	52	0,8715		-0,0946
13	0,9765	-0,0111		33	0		--	53	0		--
14	2,4903	0,6202		34	0,4609		-0,2993	54	0		--
15	0,9783	-0,0133		35	0		--	55	0,4584		-0,3063
16	0,4668	-0,4956		36	0,3718		-0,3892	56	0,8479		-0,0881
17	0	--		37	0,9004		-0,0604	57	1,776		0,3989
18	1,8157	0,2678		38	2,5055		0,2903	58	0,1232		-0,7699
19	0,7003	-0,132		39	0,5205		-0,2651	59	1,1299		0,0872
20	1,6545	0,3194		40	0,2135		-0,6657	60	2,3326		0,5292

1) the items are highlighted when the $|z| > 0.700$

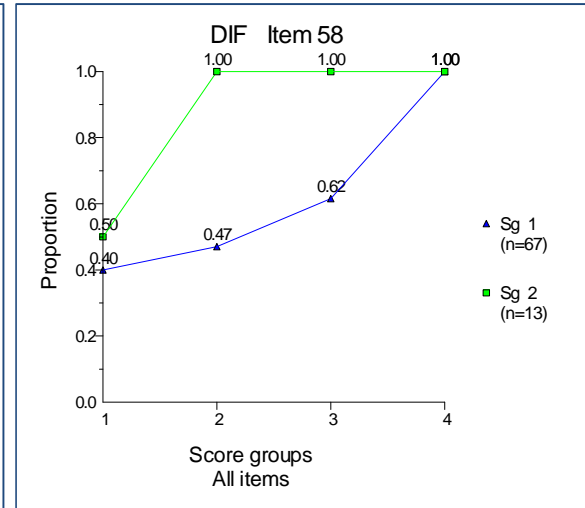
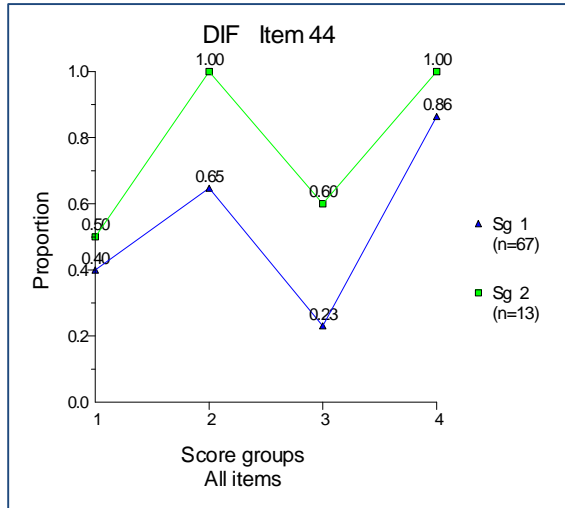
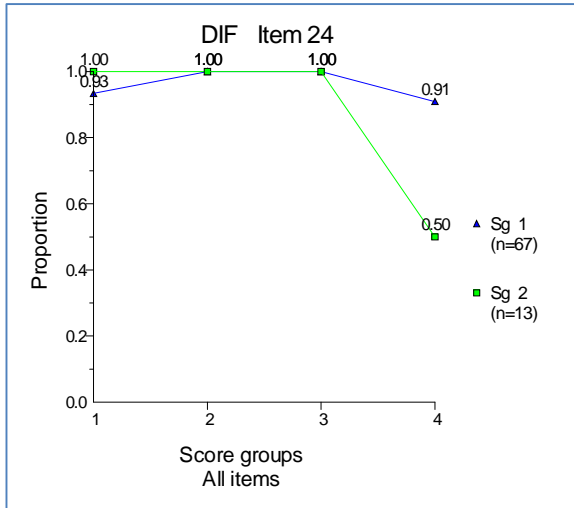


Table C.8B DIF of PCD-Lenguaje Version B

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	0	--	21	1,977	0,4402	41	1,5837	0,2749
2	0,8786	-0,0436	22	4,2857	0,4076	42	1,7543	0,2999
3	0,4176	-0,3127	23	0,862	-0,0645	43	0,4404	-0,4557
4	0,2216	-0,6833	24	1,5869	0,3331	44	1,5696	0,2914
5	4,5166	0,6344	25	0,8507	-0,0983	45	1,9911	0,3488
6	1,7508	0,3869	26	0,6748	-0,1866	46	0,696	-0,1994
7	0,9861	-0,0095	27	0,5631	-0,3761	47	2,2084	0,5385
8	0,694	-0,2318	28	2,5497	0,4419	48	3,9921	0,6971
9	0,7337	-0,1996	29	0,8955	-0,0618	49	1,6225	0,3126
10	0	--	30	0,1776	-0,6526	50	0,8279	-0,1317
11	1,3388	0,1934	31	0,9744	-0,0144	51	1,9567	0,3965
12	0,448	-0,4141	32	1,0133	0,0092	52	0,3785	-0,6226
13	0,3841	-0,5551	33	0,3566	-0,3645	53	1,9134	0,3809
14	1,3215	0,1786	34	0	--	54	3,9222	0,5255
15	0,4533	-0,4769	35	5,0954	0,5668	55	1,3385	0,1853
16	1,4521	0,177	36	2,8444	0,6611	56	2,3676	0,5035
17	1,9276	0,4365	37	0,6739	-0,1664	57	1,47	0,2505
18	1,0755	0,032	38	1,2098	0,12	58	0,4607	-0,4764
19	0,5867	-0,3154	39	0	--	59	1,6299	0,3542
20	1,5814	0,2107	40	2,3142	0,5317	60	1,8808	0,4197

1) the items are highlighted when the $|z| > 0.700$

Table C.9A DIF of PCD-Parvularia Version A

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	0	--	21	3,1302	0,5346	41	10,8624	0,8662
2	1,3291	0,1063	22	0,3165	-0,5411	42	9,9566	0,8469
3	0	--	23	0,9904	-0,0053	43	5,2317	0,6035
4	8,2409	0,9462	24	0,3847	-0,3307	44	1,2745	0,1125
5	1,7299	0,2498	25	3,7418	0,5466	45	0	--
6	0,2854	-0,4669	26	0	--	46	0	--
7	0,7744	-0,1	27	1,0427	0,0209	47	0,4332	-0,399
8	0	--	28	2,8593	0,5006	48	0,713	-0,1658
9	3,2683	0,5935	29	0,939	-0,0307	49	0,8937	-0,0548
10	0,8315	-0,0867	30	0,165	-0,6176	50	0,981	-0,0091
11	1,8703	0,2992	31	4,0883	0,6805	51	0,2717	-0,646
12	1,7322	0,2809	32	0,8009	-0,1029	52	4,7562	0,5826
13	0,2456	-0,5587	33	0,5641	-0,2885	53	3,7793	0,6745
14	0	--	34	0,3567	-0,5378	54	0,3089	-0,4245
15	0,4009	-0,4409	35	0,3496	-0,4899	55	0	--
16	0	--	36	0	--	56	1,1897	0,0908
17	0	--	37	1,2968	0,0873	57	0	--
18	0	--	38	2,158	0,3866	58	1,9739	0,3188
19	0	--	39	9,7139	1,0325	59	0,4004	-0,4112
20	0,7362	-0,1537	40	4,0924	0,7319	60	--	--

1) the items are highlighted when the $|z| > 0.700$

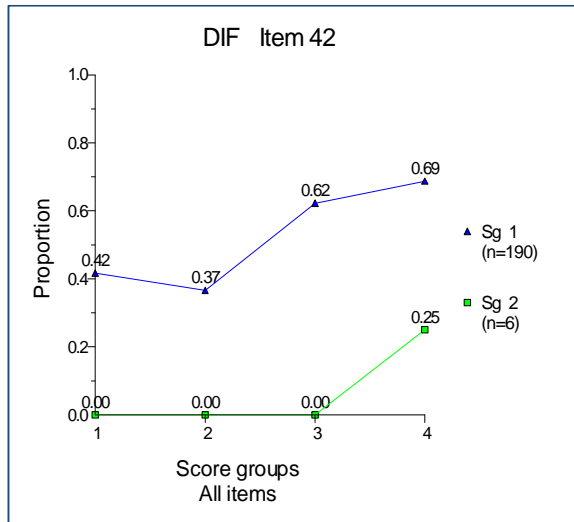
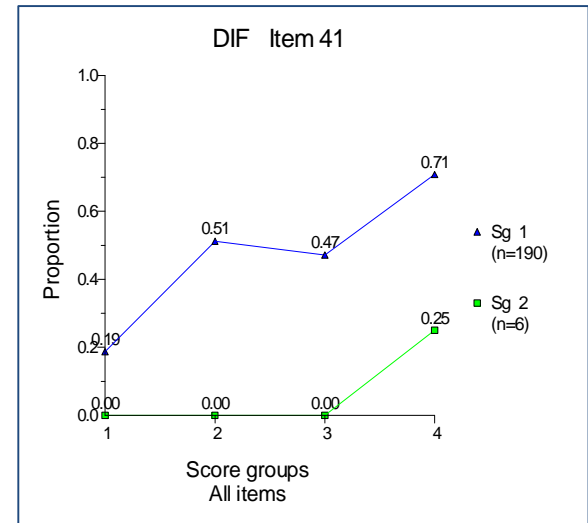
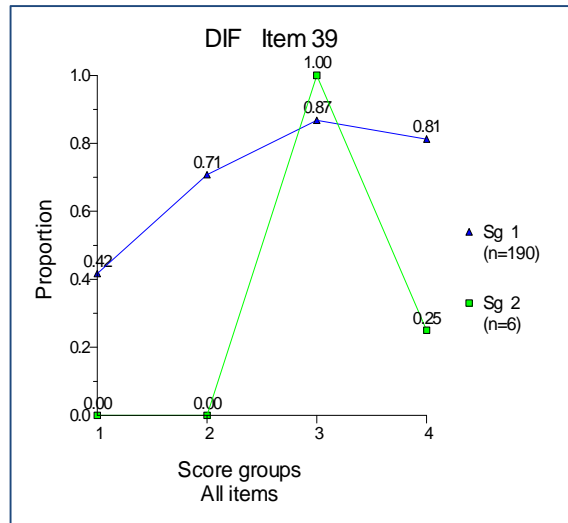
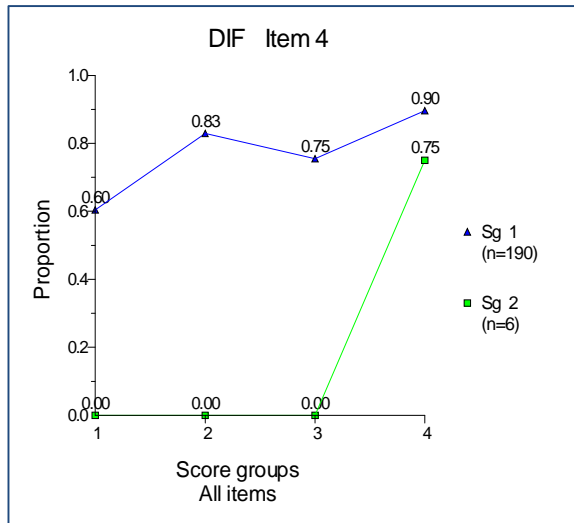


Table C.9B DIF of PCD-Parvularia Version B

Item ¹	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)	Item	DIF (MH)	stat z (standardized)
1	2,0988	0,2478	21	0,517	-0,2391	41	3,6562	0,4295
2	0	--	22	1,7439	0,2143	42	0,7083	-0,107
3	0	--	23	0	--	43	0	--
4	44,2778	1,0069	24	1,8799	0,2227	44	1,981	0,26
5	0	--	25	4,3579	0,5188	45	0	--
6	3,2727	0,4409	26	0	--	46	3,7937	0,4774
7	0,5152	-0,2209	27	6,5455	0,6246	47	0	--
8	0	--	28	0	--	48	0	--
9	1,2986	0,0847	29	3,3268	0,3993	49	2,2576	0,2766
10	3,716	0,4684	30	3,4201	0,3999	50	0	--
11	1,4596	0,1221	31	1,9414	0,2141	51	0,2774	-0,4368
12	3,0205	0,3777	32	0	--	52	--	--
13	0	--	33	--	--	53	0	--
14	0,3976	-0,3225	34	0,2174	-0,5558	54	0	--
15	2,5581	0,3301	35	0,2336	-0,4905	55	1,167	0,0526
16	0,7953	-0,0799	36	2,8855	0,3662	56	--	--
17	0	--	37	0	--	57	--	--
18	2,6984	0,3188	38	0	--	58	0	--
19	3,8722	0,4161	39	0	--	59	1,2315	0,072
20	0	--	40	0	--	60	0,2146	-0,5381

1) the items are highlighted when the $|z| > 0.700$

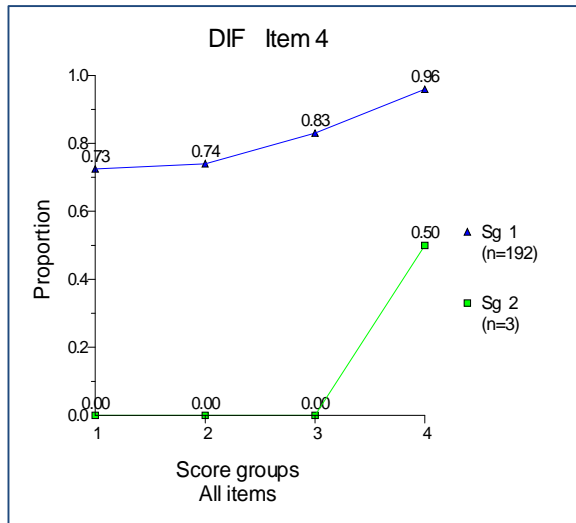


Table C.10A DIF of PCP-Parvularia Version A

Item ¹	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z
1	0	--		21	0,5936	-0,1797		41	0,3663	-0,3554	
2	0	--		22	0	--		42	0	--	
3	0,268	-0,6318		23	1,8226	0,2301		43	1,7117	0,2105	
4	3,0277	0,3802		24	0	--		44	--	--	
5	0	--		25	0,3851	-0,3398		45	0	--	
6	1,2358	0,0791		26	--	--		46	0	--	
7	0	--		27	0	--		47	1,7043	0,1916	
8	2,7056	0,3467		28	0,4591	-0,2804		48	1,2437	0,0803	
9	0,9918	-0,0028		29	2,8712	0,389		49	2,098	0,259	
10	0,262	-0,446		30	1,7399	0,1936		50	8,8494	0,901	
11	0	--		31	0,4346	-0,3542					
12	4,8074	0,6738		32	1,3792	0,123					
13	3,4963	0,4731		33	0,9828	-0,0061					
14	1,4081	0,1124		34	4,5288	0,547					
15	4,0271	0,5718		35	--	--					
16	0	--		36	2,4189	0,3567					
17	0,3839	-0,406		37	1,1517	0,0713					
18	3,3664	0,4167		38	0	--					
19	2,3008	0,3542		39	0	--					
20	1,0009	0,0003		40	2,3797	0,3623					

1) the items are highlighted when the $|z| > 0.700$

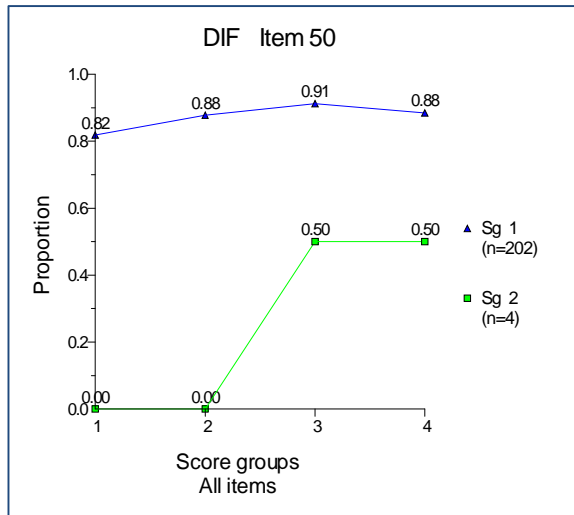


Table C.10B DIF of PCP-Parvularia Version B

Item ¹	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z
1	3,9876	0,6046		21	2,0336	0,2619		41	0,2314	-0,5238	
2	2,5943	0,3867		22	0,4235	-0,3196		42	0	--	
3	1,279	0,1159		23	0,7455	-0,1369		43	0	--	
4	0,3552	-0,3681		24	0,3205	-0,5315		44	1,6394	0,2431	
5	0	--		25	0,4804	-0,2786		45	0	--	
6	1,7949	0,2263		26	1,1297	0,0442		46	0,7805	-0,0894	
7	1,3912	0,1429		27	0	--		47	0,602	-0,1823	
8	0,3112	-0,4965		28	0	--		48	0,2492	-0,5174	
9	1,724	0,1949		29	6,1995	0,6538		49	4,2937	0,6429	
10	--	--		30	0,3935	-0,4435		50	3,373	0,566	
11	0	--		31	2,2917	0,3069					
12	0,2426	-0,5114		32	0,2618	-0,6271					
13	0	--		33	1,5206	0,197					
14	1,9807	0,2487		34	0,7839	-0,0911					
15	0,4528	-0,3696		35	0,8456	-0,0788					
16	1,611	0,2208		36	0,7457	-0,1268					
17	--	--		37	0,7392	-0,111					
18	6,4191	0,8026		38	1,0597	0,0265					
19	0	--		39	1,62	0,2061					
20	1,7141	0,2456		40	0	--					

1) the items are highlighted when the $|z| > 0.700$

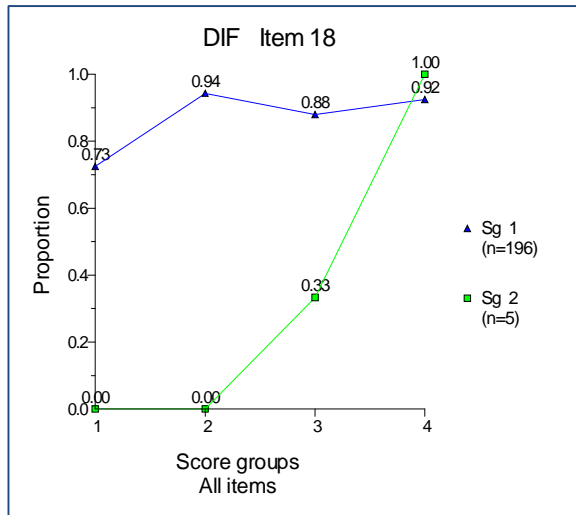


Table C.11a DIF of PCP-Media Version A

Item ¹	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z	Item	DIF (MH)	stat (standardized)	z
1	0,9153	-0,1424		21	1,034	0,0563		41	0,4434	-1,004	
2	0,817	-0,1854		22	2,094	1,0469		42	0,77	-0,3865	
3	0,902	-0,1694		23	0,9734	-0,0432		43	1,3042	0,4429	
4	1,0054	0,0091		24	0,8517	-0,1704		44	0,8891	-0,1383	
5	0,7799	-0,4206		25	1,0432	0,0564		45	1,4756	0,6141	
6	0,9046	-0,1703		26	1,3909	0,3608		46	1,5089	0,6001	
7	0,9655	-0,0504		27	1,0996	0,1567		47	0,7672	-0,4254	
8	0,9997	0		28	1,2247	0,2585		48	1,3178	0,392	
9	0,6684	-0,5706		29	0,9436	-0,0991		49	1,5102	0,6985	
10	1,2972	0,4111		30	1,3409	0,4998		50	1,025	0,0424	
11	0,5644	-0,6577		31	0,9402	-0,101					
12	1,2917	0,3582		32	1,2042	0,2995					
13	0,6856	-0,2527		33	1,706	0,7579					
14	1,0678	0,1014		34	1,1156	0,1826					
15	0,4893	-1,1526		35	1,5176	0,7016					
16	0,7992	-0,3645		36	1,4318	0,5699					
17	0,8965	-0,1569		37	0,9162	-0,1429					
18	0,965	-0,0553		38	0,9502	-0,0845					
19	0,8043	-0,3202		39	0,3677	-1,4713					
20	0,5992	-0,3933		40	0,6584	-0,6142					

1) the items are highlighted when the $|z| > 0.700$

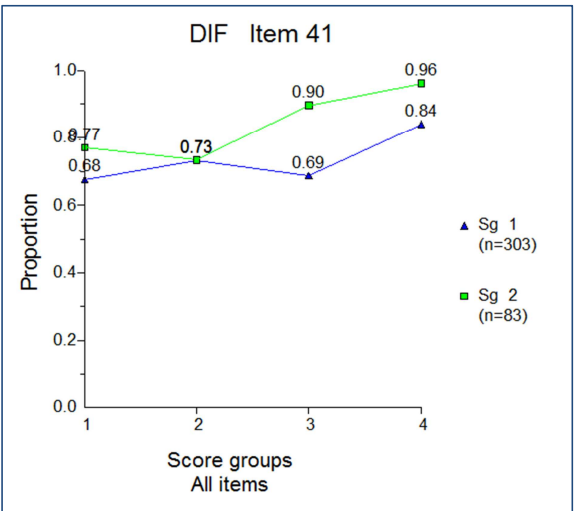
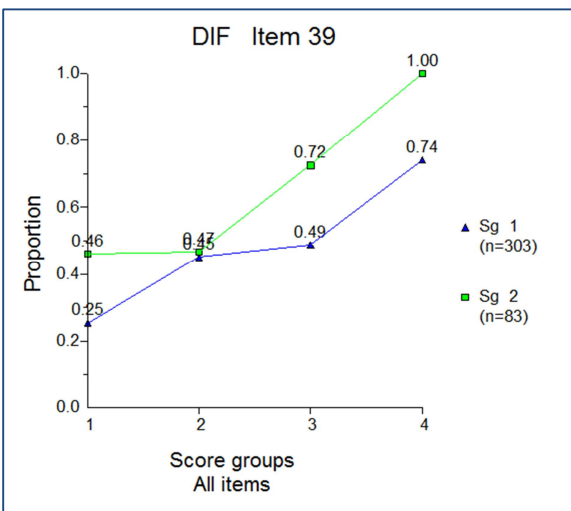
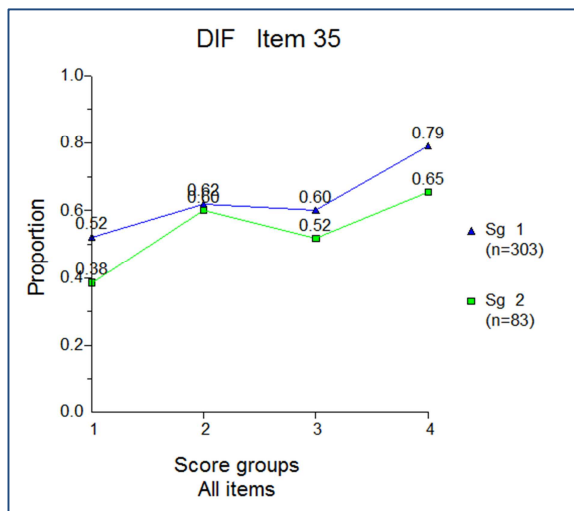
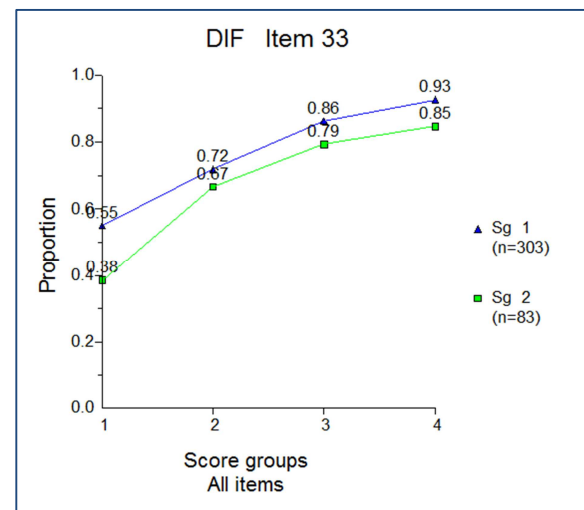
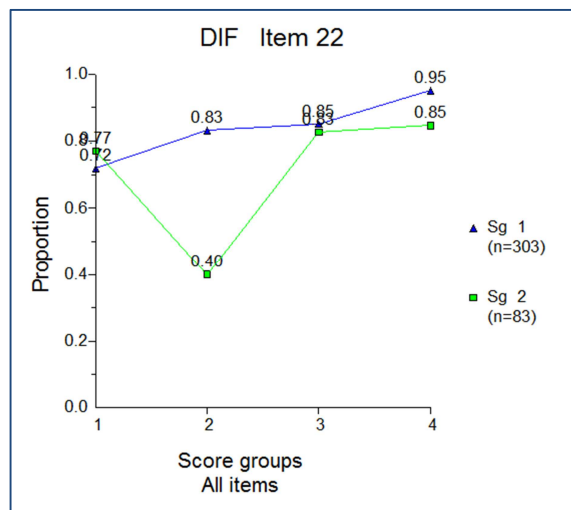
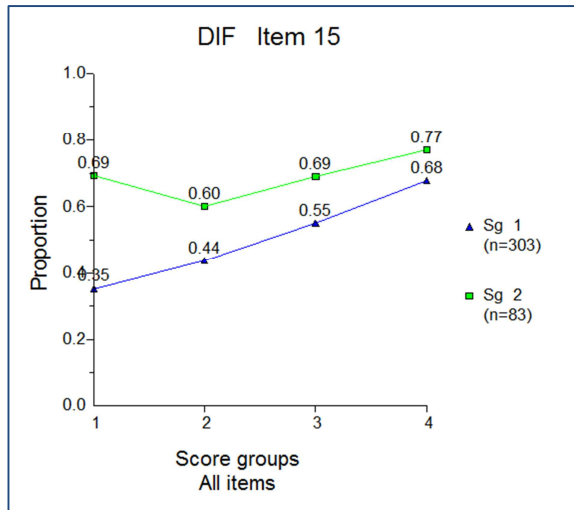
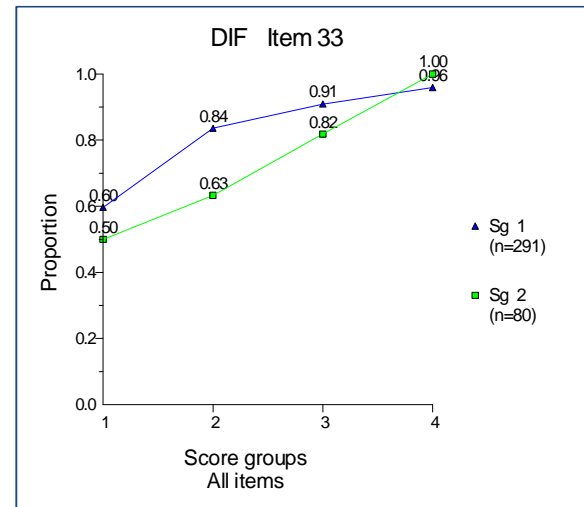
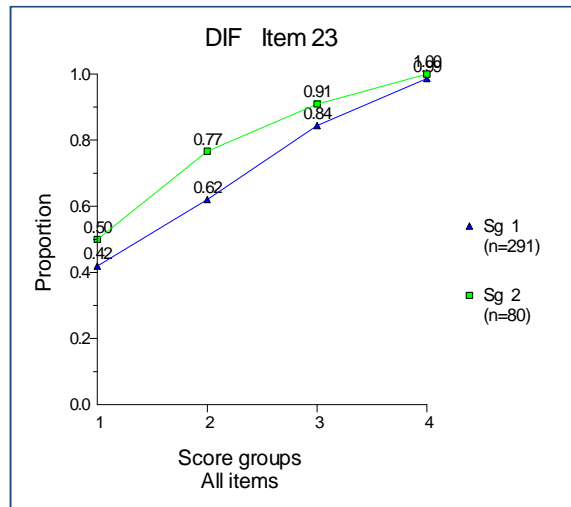
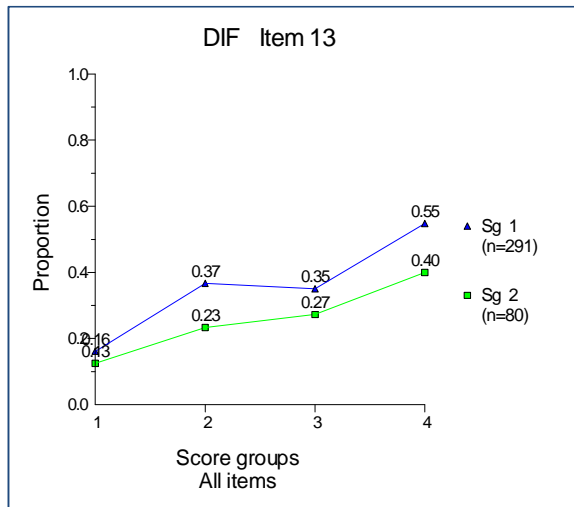
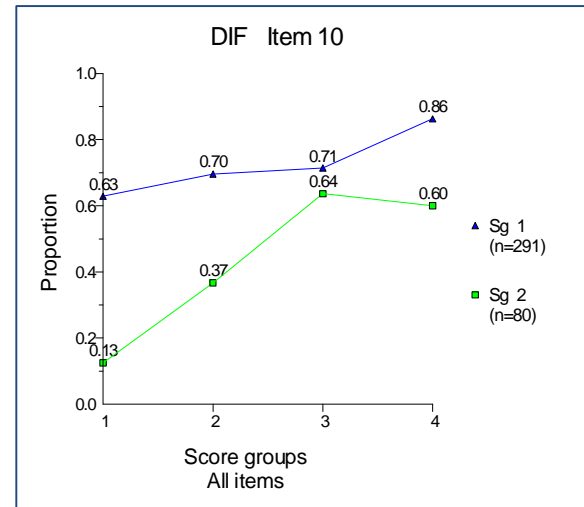
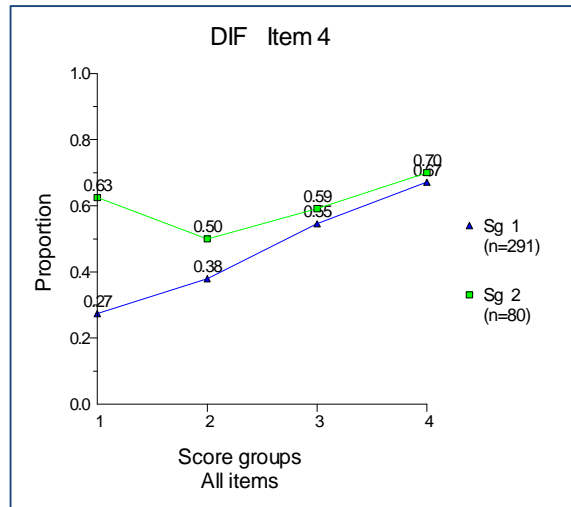
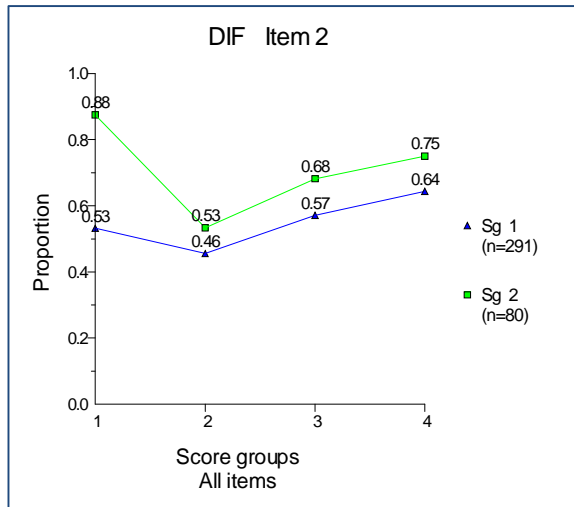
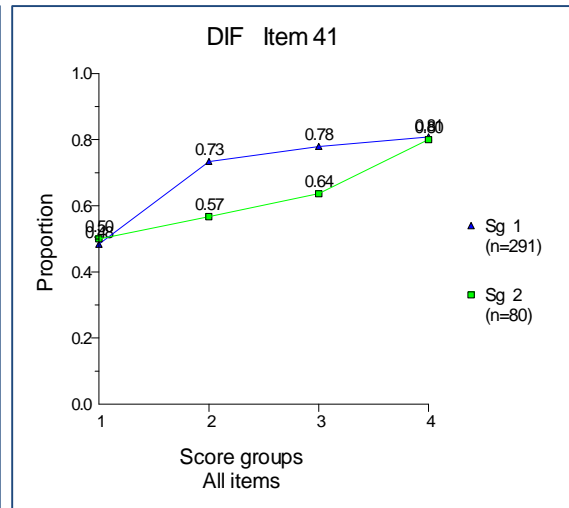
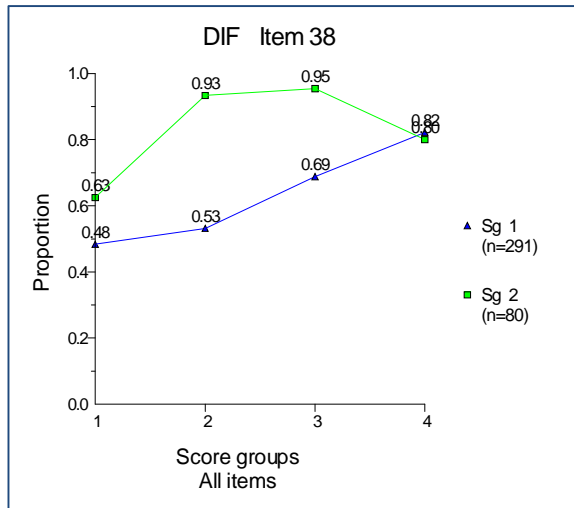


Table C.11B DIF of PCP-Media Version B

Item ¹	DIF (MH)	stat (standardized) z	Item	DIF (MH)	stat (standardized) z	Item	DIF (MH)	stat (standardized) z
1	0,9404	-0,0692	21	0,8094	-0,2718	41	1,6153	0,7545
2	0,5832	-0,8623	22	0,9394	-0,0833	42	1,5533	0,3331
3	1,0623	0,0936	23	0,5392	-0,7382	43	0,7844	-0,3877
4	0,6516	-0,709	24	1,1628	0,2123	44	0,7438	-0,3233
5	0,8811	-0,2128	25	0,5936	-0,4691	45	1,6681	0,4522
6	0,944	-0,0936	26	0,759	-0,3003	46	0,7644	-0,3404
7	1,4405	0,5424	27	1,4099	0,5511	47	1,076	0,1131
8	0,7061	-0,5225	28	0,6236	-0,5185	48	1,5231	0,6748
9	0,7227	-0,4003	29	1,339	0,488	49	1,1228	0,1686
10	3,2649	1,8917	30	1,1239	0,1944	50	1,5112	0,6293
11	0,8974	-0,1744	31	0,6946	-0,5577			
12	0,7267	-0,3594	32	0,8457	-0,1647			
13	1,6944	0,7996	33	2,0674	0,9322			
14	1,4376	0,5678	34	1,2571	0,2747			
15	0,647	-0,2845	35	0,9217	-0,1193			
16	1,2566	0,3331	36	0,9167	-0,1404			
17	1,296	0,4253	37	0,8771	-0,2166			
18	0,8726	-0,211	38	0,2706	-1,5928			
19	0,8656	-0,2195	39	0,6964	-0,4266			
20	0,6631	-0,6502	40	0,9027	-0,1577			

1) the items are highlighted when the $|z| > 0.700$





Appendix D

Equated scores in *INICÍA*

Table D.1 Equated scores in PCE-INICÍA

score	freq.	theta	SE(th)	Benchmark (± 1.5 sdt units)	INICÍA benchmarks
0	0	-3,785	1,29		
1	0	-3,163	0,809		
2	0	-2,69	0,661		
3	1	-2,308	0,562		
4	1	-2,045	0,49		
5	0	-1,852	0,442		
6	0	-1,693	0,412		
7	0	-1,551	0,394	Exceptionally low	
8	0	-1,418	0,384		
9	3	-1,285	0,38		
10	2	-1,148	0,38		
11	3	-1,006	0,382		
12	15	-0,857	0,385		
13	14	-0,706	0,387		
14	12	-0,553	0,388		
15	21	-0,401	0,387		
16	32	-0,249	0,386		
17	21	-0,097	0,384		not passed
18	49	0,053	0,381	Mediocre	Passed
19	23	0,198	0,377		
20	33	0,339	0,372		
21	24	0,473	0,368		
22	28	0,602	0,365		
23	15	0,729	0,365		
24	16	0,856	0,367		
25	12	0,986	0,371		
26	6	1,123	0,378		
27	4	1,267	0,386		

28	1	1,418	0,394		
29	1	1,574	0,403	Exceptionally high	
30	1	1,734	0,415		
31	0	1,899	0,432		
32	1	2,075	0,46		
33	0	2,275	0,506		
34	0	2,523	0,587		
35	0	2,879	0,757		
36	0	3,6	1,375		

Table D.2 Equated scores in *PCD-Basica* Versions A and B

Version A						Version B					
score	freq	theta	SE(th)	Benchmark (± 1.5 units)	INICÍA sdt benchmarks	score	freq.	theta	SE(th)	Benchmark (± 1.5 units)	INICÍA sdt benchmarks
0	0	-5,782	1,76			0	0	-5,876	1,761		
1	0	-4,659	0,903			1	0	-4,751	0,904		
2	0	-4,123	0,686			2	0	-4,214	0,687		
3	0	-3,762	0,579			3	0	-3,852	0,58		
4	0	-3,486	0,512			4	0	-3,574	0,513		
5	0	-3,26	0,466			5	0	-3,347	0,467		
6	0	-3,068	0,432			6	0	-3,154	0,433		
7	0	-2,9	0,406			7	0	-2,985	0,407		
8	0	-2,749	0,385			8	0	-2,833	0,385		
9	0	-2,612	0,367			9	0	-2,696	0,368		
10	0	-2,486	0,353			10	0	-2,57	0,353		
11	0	-2,369	0,34			11	0	-2,452	0,341		
12	0	-2,259	0,33			12	0	-2,342	0,33		
13	0	-2,156	0,32			13	0	-2,239	0,32		
14	0	-2,058	0,312			14	0	-2,141	0,312		
15	0	-1,964	0,305			15	0	-2,048	0,305		
16	0	-1,875	0,299			16	0	-1,958	0,298		
17	0	-1,788	0,293			17	0	-1,873	0,292		
18	0	-1,705	0,288			18	0	-1,79	0,287		
19	0	-1,625	0,284			19	0	-1,71	0,282		
20	1	-1,547	0,279	Exceptionally low		20	0	-1,633	0,278		

21	0	-1,471	0,276			21	0	-1,558	0,274	Exceptionally low	
22	2	-1,396	0,272			22	0	-1,485	0,27		
23	1	-1,324	0,269			23	3	-1,414	0,267		
24	1	-1,253	0,267			24	0	-1,344	0,264		
25	0	-1,183	0,264			25	2	-1,276	0,261		
26	0	-1,115	0,262			26	1	-1,209	0,259		
27	3	-1,048	0,26			27	2	-1,143	0,256		
28	9	-0,981	0,258			28	2	-1,079	0,254		
29	2	-0,916	0,256			29	4	-1,015	0,253		
30	4	-0,851	0,255			30	4	-0,952	0,251		
31	4	-0,787	0,254			31	5	-0,89	0,249		
32	1	-0,724	0,252			32	4	-0,829	0,248		
33	2	-0,661	0,251			33	7	-0,768	0,247		
34	7	-0,598	0,251			34	4	-0,708	0,246		
35	6	-0,536	0,25			35	9	-0,648	0,245		
36	10	-0,474	0,249			36	7	-0,589	0,244		
37	9	-0,413	0,249			37	9	-0,53	0,244		
38	10	-0,352	0,249			38	6	-0,471	0,243		
39	16	-0,29	0,248			39	14	-0,413	0,243		
Continuing...											
40	10	-0,229	0,248			40	6	-0,354	0,243		
41	14	-0,168	0,248			41	13	-0,296	0,243		
42	9	-0,107	0,248			42	4	-0,237	0,243		
43	12	-0,046	0,249			43	17	-0,179	0,243		
44	11	0,016	0,249	Mediocre		44	9	-0,12	0,243		
45	13	0,078	0,25			45	22	-0,061	0,244		
46	21	0,14	0,251		insufficient	46	5	-0,002	0,245		insufficient
47	17	0,202	0,251		sufficient	47	6	0,057	0,245	Mediocre	sufficient

48	17	0,265	0,252		48	8	0,117	0,246		
49	15	0,328	0,253		49	13	0,177	0,247		
50	19	0,392	0,255		50	18	0,238	0,249		
51	12	0,457	0,256		51	17	0,3	0,25		
52	6	0,522	0,258		52	13	0,362	0,252		
53	9	0,589	0,26		53	10	0,425	0,253		
54	6	0,656	0,262		54	11	0,489	0,255		
55	11	0,724	0,264		55	8	0,554	0,258		
56	10	0,794	0,267		56	10	0,62	0,26		
57	7	0,865	0,269		57	9	0,688	0,263		
58	9	0,938	0,273	sufficient	58	5	0,757	0,266		sufficient
59	6	1,012	0,276	exceptional	59	6	0,828	0,269		exceptional
60	2	1,088	0,28		60	6	0,9	0,273		
61	4	1,167	0,284		61	7	0,975	0,277		
62	1	1,247	0,289		62	4	1,052	0,282		
63	3	1,331	0,294		63	0	1,132	0,287		
64	2	1,418	0,3		64	2	1,214	0,293		
65	1	1,508	0,306	Exceptionally high	65	6	1,3	0,299		
66	3	1,602	0,314		66	2	1,39	0,307		
67	0	1,701	0,322		67	2	1,485	0,315		
68	0	1,805	0,331		68	0	1,584	0,324	Exceptionally high	
69	0	1,916	0,342		69	0	1,69	0,335		
70	0	2,035	0,355		70	1	1,804	0,347		
71	1	2,162	0,37		71	0	1,926	0,362		
72	0	2,301	0,387		72	1	2,059	0,38		
73	0	2,454	0,409		73	0	2,206	0,401		
74	0	2,625	0,435		74	0	2,371	0,428		
75	0	2,821	0,47		75	0	2,559	0,462		

76	0	3,05	0,516			76	0	2,781	0,508		
77	0	3,331	0,583			77	0	3,053	0,575		
78	0	3,697	0,691			78	0	3,41	0,683		
79	0	4,239	0,907			79	0	3,942	0,9		
80	0	5,37	1,767			80	0	5,061	1,755		

Table D.3 Equated scores in *PCP-Basica* Versions A and B

Version A				Version B									
score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICIA benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICIA benchmarks
0	0	-5,66	1,78				0	0	-5,699	1,775			
1	0	-4,517	0,917				1	0	-4,562	0,914			
2	0	-3,963	0,701				2	0	-4,012	0,698			
3	0	-3,584	0,595				3	0	-3,637	0,592			
4	0	-3,291	0,529				4	0	-3,347	0,527			
5	0	-3,049	0,484				5	0	-3,107	0,482			
6	0	-2,84	0,451				6	0	-2,901	0,449			
7	0	-2,656	0,425				7	0	-2,718	0,424			
8	0	-2,49	0,405				8	0	-2,553	0,403			
9	0	-2,338	0,388				9	0	-2,402	0,387			
10	0	-2,197	0,374				10	0	-2,261	0,374			
11	0	-2,064	0,363				11	0	-2,129	0,362			
12	0	-1,94	0,353				12	0	-2,005	0,353			
13	0	-1,821	0,344				13	0	-1,886	0,345			

14	0	-	1,707	0,337		14	0	-	1,772	0,338	
15	0	-	1,598	0,331	Exceptionally low	15	1	-	1,662	0,332	
16	1	-	1,492	0,326		16	0	-	1,556	0,327	Exceptionally low
17	2	-	1,389	0,321		17	0	-	1,452	0,322	
18	2	-	1,289	0,317		18	1	-	1,351	0,319	
19	0	-	1,191	0,314		19	2	-	1,252	0,316	
20	2	-	1,095	0,312		20	2	-	1,155	0,313	
21	4	-	1	0,31		21	4	-	1,059	0,311	
22	5	-	0,906	0,308		22	8	-	0,964	0,31	
23	5	-	0,813	0,307		23	5	-	-0,87	0,309	
24	10	-	0,721	0,306		24	8	-	0,776	0,308	
25	8	-	0,628	0,306		25	11	-	0,683	0,308	
26	6	-	0,536	0,306		26	7	-	0,589	0,308	
27	8	-	0,444	0,307		27	13	-	0,495	0,309	
28	21	-	0,351	0,308		28	20	-	0,401	0,31	
29	11	-	0,257	0,31		29	17	-	0,305	0,312	
30	17	-		0,312	insufficient	30	24	-		0,314	insufficient

		0,162						0,209			
31	25	-0,066	0,314		sufficient	31	24	-0,111	0,316		sufficient
32	23	0,032	0,317	Mediocre		32	30	-0,012	0,319		
33	19	0,132	0,321			33	19	0,09	0,323	Mediocre	
34	29	0,235	0,326			34	30	0,194	0,328		
35	26	0,341	0,331			35	18	0,301	0,333		
36	21	0,45	0,337			36	24	0,411	0,339		
37	22	0,563	0,344			37	19	0,526	0,346		
38	17	0,682	0,353		sufficient	38	17	0,646	0,354		sufficient
39	15	0,807	0,362		exceptional	39	9	0,771	0,363		exceptional
Continuing...											
40	6	0,939	0,374			40	12	0,904	0,374		
41	9	1,08	0,388			41	8	1,045	0,388		
42	5	1,232	0,405			42	6	1,197	0,404		
43	3	1,398	0,425			43	1	1,362	0,424		
44	1	1,583	0,451	Exceptionally High		44	3	1,545	0,449	Exceptionally High	
45	1	1,792	0,485			45	2	1,751	0,482		
46	0	2,036	0,531			46	0	1,991	0,527		
47	0	2,331	0,597			47	0	2,281	0,592		
48	0	2,714	0,704			48	0	2,656	0,698		
49	0	3,273	0,921			49	0	3,205	0,914		
50	0	4,424	1,787			50	0	4,342	1,774		

Table D.4 Equated scores in *PCD-Biología*

score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICÍA benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICÍA benchmarks
0	0	-5,79	1,812										
1	0	-4,609	0,935				31	6	-0,138	0,293			
2	0	-4,028	0,716				32	3	-0,053	0,294			
3	0	-3,63	0,607				33	5	0,032	0,295	Mediocre		
4	0	-3,323	0,539				34	7	0,118	0,296			
5	0	-3,071	0,492				35	6	0,205	0,297			insufficient
6	0	-2,855	0,458				36	7	0,293	0,299			sufficient
7	0	-2,665	0,431				37	5	0,382	0,301			
8	0	-2,494	0,409				38	1	0,472	0,304			
9	0	-2,339	0,392				39	1	0,564	0,307			
10	0	-2,195	0,377				40	5	0,658	0,31			sufficient
11	0	-2,061	0,365				41	1	0,754	0,314			exceptional
12	0	-1,935	0,354				42	2	0,853	0,319			
13	0	-1,815	0,345				43	2	0,955	0,324			
14	0	-1,701	0,337				44	2	1,06	0,33			
15	1	-	0,33	Exceptionally			45	3	1,169	0,336			

		1,592	low						
16	0	-1,487	0,324	46	1	1,282	0,344		
17	2	-1,385	0,319	47	0	1,401	0,353		
18	0	-1,286	0,314	48	0	1,526	0,363	Exceptionally low	
19	0	-1,19	0,31	49	1	1,659	0,374		
20	1	-1,096	0,307	50	0	1,801	0,388		
21	0	-1,004	0,304	51	0	1,954	0,404		
22	0	-0,913	0,301	52	0	2,121	0,423		
23	2	-0,824	0,299	53	0	2,304	0,446		
24	3	-0,736	0,297	54	0	2,51	0,475		
25	3	-0,65	0,296	55	0	2,745	0,512		
26	0	-0,563	0,295	56	0	3,021	0,562		
27	2	-0,478	0,294	57	0	3,357	0,631		
28	4	-0,393	0,293	58	0	3,79	0,741		
29	2	-0,308	0,293	59	0	4,41	0,961		
30	2	-0,223	0,293	60	0	5,63	1,849		

Table D.5 Equated scores in *PCD-Física*

score	freq.	theta	SE(th)	Benchmark (± 1.5 sdt units)	INICÍA benchmarks	score	freq.	theta	SE(th)	Benchmark (± 1.5 sdt units)	INICÍA benchmarks
0	0	-5,302	1,793								
1	0	-4,144	0,924			31	1	0,073	0,277	Mediocre	
2	0	-3,578	0,706			32	1	0,148	0,278		
3	0	-3,192	0,598			33	3	0,225	0,278		
4	0	-2,895	0,531			34	1	0,301	0,279		
5	0	-2,65	0,484			35	0	0,378	0,28		
6	0	-2,442	0,449			36	2	0,456	0,282		
7	0	-2,26	0,422			37	1	0,535	0,284		insufficient
8	0	-2,096	0,4			38	2	0,615	0,286		sufficient
9	1	-1,948	0,382			39	0	0,697	0,289		
10	0	-1,811	0,367			40	0	0,78	0,292		
11	0	-1,685	0,354			41	0	0,865	0,295		
12	1	-1,566	0,343	Exceptional low		42	2	0,952	0,299		
13	1	-1,454	0,334			43	1	1,041	0,304		
14	0	-1,347	0,325			44	4	1,133	0,309		
15	0	-	0,318			45	0	1,229	0,315		

			1,246							
16	2	-	0,312	46	1	1,328	0,322			
17	2	-	0,306	47	1	1,432	0,329			
18	2	-	0,301	48	0	1,541	0,338	Exceptional	high	
19	1	-	0,297	49	1	1,656	0,348			
20	4	-	0,293	50	1	1,778	0,36			
21	0	-	0,29	51	1	1,909	0,375			sufficient
22	2	-	0,287	52	0	2,051	0,392			exceptional
23	2	-	0,285	53	1	2,207	0,412			
24	2	-	0,283	54	0	2,381	0,438			
25	1	-	0,281	55	0	2,578	0,472			
26	2	-	0,28	56	1	2,809	0,518			
27	0	-	0,279	57	0	3,09	0,584			
28	0	-	0,278	58	0	3,457	0,691			
29	3	-	0,277	59	0	3,997	0,907			
30	3	-	0,277	60	0	5,126	1,765			

Table D.6 Equated scores in *PCD-Matemática*

score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICÍA benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICÍA benchmarks
0	0	-5,817	1,783										
1	0	-4,671	0,919				31	5	-0,374	0,284			
2	0	-4,114	0,703				32	9	-0,294	0,284			
3	0	-3,732	0,596				33	5	-0,214	0,285			
4	0	-3,437	0,53				34	9	-0,134	0,286			
5	0	-3,193	0,485				35	5	-0,052	0,287			
6	0	-2,984	0,451				36	8	0,03	0,289	Mediocre		
7	0	-2,8	0,424				37	6	0,112	0,291			
8	0	-2,634	0,403				38	5	0,196	0,293			insufficient
9	0	-2,483	0,386				39	8	0,282	0,296			sufficient
10	0	-2,343	0,372				40	3	0,369	0,299			
11	0	-2,212	0,359				41	8	0,458	0,302			
12	0	-2,09	0,349				42	9	0,549	0,306			
13	0	-1,974	0,339				43	3	0,643	0,311			
14	1	-1,863	0,331				44	5	0,739	0,316			
15	0	-1,758	0,325				45	3	0,839	0,322			
16	0	-1,656	0,318				46	4	0,943	0,329			
17	0	-1,558	0,313	Exceptionally low			47	4	1,051	0,336			
18	2	-1,463	0,308				48	6	1,165	0,345			
19	1	-1,371	0,304				49	4	1,285	0,356			
20	1	-1,281	0,3				50	4	1,413	0,368			sufficient

21	1	-1,193	0,297	51	4	1,55	0,382	Exceptionally high	exceptional
22	4	-1,107	0,294	52	5	1,698	0,399		
23	3	-1,022	0,292	53	5	1,86	0,42		
24	8	-0,938	0,29	54	1	2,04	0,446		
25	6	-0,856	0,288	55	2	2,244	0,48		
26	0	-0,775	0,287	56	1	2,482	0,525		
27	1	-0,694	0,286	57	1	2,772	0,591		
28	5	-0,613	0,285	58	0	3,146	0,698		
29	10	-0,533	0,284	59	0	3,696	0,914		
30	4	-0,454	0,284	60	0	4,834	1,775		

Table D.7 Equated scores in *PCD-Química*

score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks
0	0	-5,877	1,813								
1	0	-4,696	0,937			31	2	-0,184	0,291		
2	0	-4,113	0,718			32	1	-0,1	0,291		
3	0	-3,713	0,609			33	2	-0,016	0,292		
4	0	-3,403	0,542			34	4	0,068	0,292	Mediocre	
5	0	-3,148	0,495			35	3	0,153	0,293		
6	0	-2,929	0,461			36	2	0,238	0,295		
7	0	-2,737	0,434			37	2	0,325	0,297		
8	0	-2,564	0,412			38	1	0,412	0,299		insufficient
9	0	-2,405	0,395			39	3	0,501	0,301		sufficient
10	0	-2,259	0,38			40	1	0,591	0,304		
11	0	-2,122	0,368			41	1	0,683	0,308		
12	0	-1,994	0,357			42	0	0,778	0,312		
13	0	-1,872	0,348			43	0	0,875	0,316		
14	0	-	0,34			44	2	0,975	0,321		

		1,756								
15	0	-1,644	0,333		45	1	1,078	0,327		
16	0	-1,537	0,327	Exceptionally low	46	1	1,185	0,334		
17	0	-1,434	0,321		47	0	1,298	0,342		
18	0	-1,333	0,317		48	0	1,415	0,351		
19	1	-1,236	0,312		49	0	1,539	0,362	Exceptionally high	
20	0	-1,14	0,309		50	2	1,671	0,374		sufficient
21	0	-1,047	0,305		51	1	1,813	0,388		exceptional
22	1	-0,956	0,302		52	0	1,966	0,406		
23	0	-0,866	0,3		53	0	2,134	0,427		
24	0	-0,778	0,298		54	0	2,32	0,453		
25	2	-0,691	0,296		55	0	2,531	0,487		
26	2	-0,605	0,294		56	0	2,777	0,533		
27	1	-0,52	0,293		57	0	3,076	0,599		
28	3	-0,435	0,292		58	0	3,461	0,706		
29	3	-0,351	0,291		59	0	4,022	0,922		
30	1	-0,267	0,291		60	0	5,174	1,788		

Table D.8 Equated scores in *PCD-Historia* Versions A and B

Version A				Version B							
score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks
0	0	-5,919	1,769			0	0	-5,957	1,77		
1	0	-4,786	0,91			1	0	-4,825	0,91		
2	0	-4,242	0,693			2	0	-4,28	0,694		
3	0	-3,873	0,586			3	0	-3,91	0,588		
4	0	-3,589	0,52			4	0	-3,625	0,522		
5	0	-3,355	0,474			5	0	-3,39	0,477		
6	0	-3,156	0,441			6	0	-3,188	0,443		
7	0	-2,981	0,414			7	0	-3,01	0,417		
8	0	-2,823	0,394			8	0	-2,85	0,397		
9	0	-2,68	0,377			9	0	-2,704	0,38		
10	0	-2,547	0,362			10	0	-2,569	0,366		
11	0	-2,423	0,35			11	0	-2,442	0,354		
12	0	-2,307	0,34			12	0	-2,323	0,344		
13	0	-2,196	0,331			13	0	-2,21	0,335		
14	0	-2,091	0,324			14	0	-2,103	0,328		
15	0	-1,99	0,317			15	0	-1,999	0,321		
16	0	-1,894	0,311			16	0	-1,9	0,315		

17	0	-1,8	0,306		17	0	-1,804	0,31		
18	0	-1,709	0,302		18	0	-1,71	0,306		
19	0	-1,62	0,298		19	0	-1,62	0,302		
20	0	-1,534	0,295	Exceptionally low	20	0	-1,531	0,298	Exceptionally low	
21	0	-1,449	0,292		21	0	-1,444	0,295		
22	0	-1,366	0,289		22	0	-1,359	0,292		
23	1	-1,284	0,287		23	0	-1,276	0,29		
24	1	-1,204	0,285		24	1	-1,193	0,288		
25	0	-1,124	0,284		25	1	-1,112	0,286		
26	3	-1,045	0,282		26	0	-1,031	0,285		
27	0	-0,966	0,282		27	1	-0,951	0,284		
28	5	-0,888	0,281		28	1	-0,872	0,283		
29	4	-0,81	0,281		29	4	-0,793	0,282		
30	4	-0,733	0,281		30	4	-0,714	0,282		
31	6	-0,655	0,281		31	4	-0,635	0,282		
32	4	-0,577	0,281		32	6	-0,557	0,282		
33	3	-0,498	0,282		33	2	-0,478	0,283		
34	8	-0,419	0,283		34	7	-	0,284		

59	0	3,676	0,976			59	0	3,36	0,91		
60	0	4,964	1,895			60	0	4,493	1,77		

Table D.9 Equated scores in *PCD-Lenguaje* Versions A and B

Version A				Version B							
score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks
0	0	-6,185	1,818			0	0	-6,085	1,793		
1	0	-4,998	0,939			1	0	-4,926	0,924		
2	0	-4,412	0,719			2	0	-4,361	0,706		
3	0	-4,01	0,61			3	0	-3,974	0,598		
4	0	-3,699	0,542			4	0	-3,677	0,531		
5	0	-3,444	0,495			5	0	-3,432	0,485		
6	0	-3,225	0,46			6	0	-3,223	0,45		
7	0	-3,033	0,432			7	0	-3,04	0,423		
8	0	-2,861	0,41			8	0	-2,875	0,402		
9	0	-2,704	0,392			9	0	-2,725	0,384		
10	0	-2,56	0,377			10	0	-2,587	0,369		
11	0	-2,425	0,364			11	0	-2,458	0,357		
12	0	-2,299	0,353			12	0	-2,338	0,346		
13	0	-2,18	0,343			13	0	-2,224	0,337		

14	0	-	2,067	0,335		14	0	-	2,115	0,328	
15	0	-	1,959	0,328		15	0	-	2,012	0,321	
16	0	-	1,856	0,321		16	0	-	1,912	0,315	
17	0	-	1,756	0,315		17	0	-	1,816	0,31	
18	0	-	1,66	0,31		18	0	-	1,724	0,305	
19	0	-	1,566	0,306	Exceptionally low	19	0	-	1,633	0,3	Exceptionally low
20	0	-	1,476	0,302		20	0	-	1,546	0,297	
21	0	-	1,387	0,298		21	0	-	-1,46	0,293	
22	0	-	1,3	0,295		22	0	-	1,376	0,29	
23	0	-	1,215	0,292		23	0	-	1,294	0,288	
24	0	-	1,132	0,29		24	0	-	1,213	0,286	
25	0	-	1,05	0,288		25	0	-	1,133	0,284	
26	0	-	0,968	0,286		26	0	-	1,053	0,282	
27	0	-	0,888	0,285		27	0	-	0,975	0,281	
28	1	-	0,808	0,284		28	0	-	0,897	0,28	
29	1	-	0,729	0,283		29	0	-	-0,82	0,28	
30	0	-	-0,65	0,282		30	1	-	-	0,279	

51	3	1,291	0,375	exceptional	51	2	1,184	0,375	exceptional
52	0	1,433	0,392		52	4	1,326	0,392	
53	1	1,588	0,412	Exceptionally high	53	0	1,483	0,413	Exceptionally high
54	2	1,761	0,438		54	4	1,657	0,439	
55	0	1,958	0,472		55	1	1,855	0,473	
56	0	2,189	0,517		56	0	2,087	0,518	
57	0	2,469	0,583		57	0	2,369	0,585	
58	0	2,835	0,69		58	0	2,737	0,692	
59	0	3,375	0,906		59	1	3,279	0,908	
60	0	4,502	1,764		60	0	4,409	1,767	

Table D.10 Equated scores in *PCD-Parvularia* Versions A and B

Version A				Version B							
score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks
0	0	-5,565	1,766			0	0	-5,494	1,764		
1	0	-4,436	0,908			1	0	-4,366	0,906		
2	0	-3,894	0,691			2	0	-3,826	0,69		
3	0	-3,527	0,584			3	0	-3,461	0,583		
4	0	-3,245	0,518			4	0	-3,18	0,517		
5	0	-3,014	0,473			5	0	-2,95	0,471		
6	0	-2,816	0,439			6	0	-2,753	0,438		
7	0	-2,642	0,413			7	0	-2,58	0,412		
8	0	-2,485	0,392			8	0	-2,425	0,391		
9	0	-2,343	0,376			9	0	-2,283	0,374		
10	0	-2,211	0,361			10	0	-2,152	0,36		
11	0	-2,088	0,35			11	0	-2,03	0,348		
12	0	-1,972	0,34			12	0	-1,915	0,338		
13	0	-1,862	0,331			13	0	-1,807	0,329		

14	0	-	1,757	0,323		14	0	-	1,703	0,321	
15	0	-	1,657	0,317		15	0	-	1,604	0,315	
16	0	-1,56	0,311	Exceptionally low		16	0	-	1,508	0,309	Exceptionally low
17	1	-	1,467	0,306		17	1	-	1,416	0,304	
18	0	-	1,376	0,302		18	0	-	1,326	0,299	
19	0	-	1,287	0,298		19	1	-	1,239	0,295	
20	2	-	1,201	0,295		20	2	-	1,154	0,292	
21	3	-	1,116	0,292		21	2	-	1,071	0,289	
22	1	-	1,033	0,289		22	3	-	-0,99	0,286	
23	3	-	0,951	0,287		23	6	-	0,909	0,284	
24	1	-	-0,87	0,285		24	2	-	-0,83	0,282	
25	7	-	-0,79	0,284		25	8	-	0,752	0,281	
26	8	-	0,711	0,283		26	5	-	0,674	0,279	
27	4	-	0,632	0,282		27	7	-	0,598	0,279	
28	7	-	0,554	0,282		28	8	-	0,521	0,278	
29	10	-	0,475	0,281		29	1	-	0,445	0,277	
30	8	-	0,397	0,281		30	4	-	0,369	0,277	

31	11	-	0,319	0,282		31	11	-	0,293	0,277	
32	15	-	0,241	0,282		32	9	-	0,217	0,278	
33	8	-	0,162	0,283		33	17	-	0,141	0,278	
34	10	-	0,082	0,284		34	15	-	0,064	0,279	
35	8	-	0,002	0,285	insufficient	35	9	0,013	0,281	Mediocre	insufficient
36	13	0,078	0,287	Mediocre	sufficient	36	15	0,091	0,282		sufficient
37	14	0,16	0,289			37	7	0,17	0,284		
38	9	0,243	0,291			38	9	0,25	0,286		
39	11	0,328	0,294			39	7	0,332	0,289		
Continuing...											
40	11	0,414	0,297			40	11	0,415	0,292		
41	6	0,502	0,301		sufficient	41	13	0,5	0,295		sufficient
42	6	0,593	0,305		exceptional	42	7	0,587	0,299		exceptional
43	7	0,686	0,309			43	4	0,676	0,304		
44	1	0,781	0,315			44	2	0,768	0,309		
45	5	0,881	0,321			45	2	0,863	0,315		
46	2	0,984	0,327			46	4	0,962	0,321		
47	2	1,091	0,335			47	1	1,066	0,329		
48	0	1,204	0,344			48	1	1,174	0,338		
49	0	1,323	0,354			49	0	1,289	0,348		
50	0	1,45	0,366			50	0	1,411	0,36		
51	0	1,585	0,38	Exceptionally high		51	0	1,542	0,374	Exceptionally high	
52	0	1,732	0,397			52	0	1,683	0,391		
53	0	1,893	0,418			53	1	1,839	0,412		
54	0	2,071	0,444			54	0	2,012	0,438		

55	0	2,274	0,478		55	0	2,208	0,471	
56	0	2,51	0,523		56	0	2,439	0,517	
57	0	2,798	0,589		57	0	2,72	0,583	
58	0	3,171	0,696		58	0	3,086	0,69	
59	0	3,718	0,912		59	0	3,626	0,907	
60	0	4,854	1,773		60	0	4,754	1,765	

Table D.11 Equated scores in *PCP-Parvularia* Versions A and B

Version A				Version B							
score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	INICÍA sdt benchmarks
0	0	-5,42	1,771			0	0	-5,444	1,772		
1	0	-4,286	0,912			1	0	-4,31	0,912		
2	0	-3,74	0,696			2	0	-3,763	0,696		
3	0	-3,367	0,59			3	0	-3,391	0,59		
4	0	-3,08	0,524			4	0	-3,103	0,525		
5	0	-2,842	0,48			5	0	-2,865	0,48		
6	0	-2,638	0,447			6	0	-2,66	0,447		
7	0	-2,457	0,422			7	0	-2,479	0,422		
8	0	-2,294	0,402			8	0	-2,316	0,402		
9	0	-2,144	0,385			9	0	-2,166	0,386		
10	1	-2,005	0,372			10	0	-2,026	0,372		
11	0	-1,874	0,361			11	0	-1,895	0,361		
12	0	-1,75	0,352			12	0	-1,771	0,352		
13	0	-1,632	0,344			13	0	-1,652	0,344		

14	0	-1,519	0,337	Exceptionally low		14	2	-1,539	0,337	Exceptionally low	
15	2	-1,409	0,331			15	1	-1,429	0,331		
16	2	-1,304	0,326			16	1	-1,323	0,326		
17	1	-1,201	0,322			17	2	-1,22	0,322		
18	2	-1,1	0,318			18	1	-1,119	0,319		
19	2	-1,001	0,315			19	4	-1,02	0,316		
20	4	-0,904	0,313			20	3	-0,922	0,314		
21	7	-0,808	0,311			21	7	-0,826	0,312		
22	2	-0,714	0,31			22	11	-0,731	0,31		
23	7	-0,619	0,309			23	5	-0,636	0,309		
24	12	-0,526	0,308			24	6	-0,542	0,309		
25	11	-0,432	0,308			25	7	-0,448	0,309		
26	11	-0,339	0,308			26	15	-0,354	0,309		
27	15	-0,245	0,309			27	14	-0,26	0,31		
28	11	-0,15	0,31			28	9	-0,165	0,311		
29	14	-0,055	0,312			29	17	-0,069	0,313		
30	17	0,042	0,314	Mediocre	insufficient	30	11	0,029	0,315	Mediocre	insufficient

31	10	0,139	0,317	sufficient	31	12	0,127	0,318	sufficient
32	17	0,239	0,32		32	11	0,228	0,321	
33	9	0,341	0,324		33	12	0,331	0,325	
34	12	0,446	0,328		34	11	0,436	0,33	
35	11	0,553	0,334	sufficient	35	9	0,544	0,335	sufficient
36	5	0,664	0,34	exceptional	36	9	0,657	0,341	exceptional
37	8	0,78	0,347		37	4	0,773	0,349	
38	1	0,9	0,355		38	3	0,895	0,357	
39	4	1,027	0,365		39	0	1,023	0,367	
Continuing...									
40	1	1,161	0,377		40	3	1,159	0,379	
41	2	1,305	0,391		41	0	1,304	0,393	
42	0	1,459	0,408		42	3	1,46	0,409	
43	0	1,628	0,428	Exceptionally high	43	0	1,63	0,43	Exceptionally high
44	1	1,815	0,454		44	0	1,819	0,456	
45	0	2,026	0,488		45	0	2,032	0,49	
46	0	2,273	0,533		46	0	2,28	0,535	
47	0	2,571	0,599		47	0	2,581	0,601	
48	0	2,957	0,707		48	0	2,968	0,708	
49	0	3,52	0,923		49	0	3,533	0,925	
50	0	4,675	1,79		50	0	4,689	1,792	

Table D.12 Equated scores in *PCP-Media* Versions A and B

Version A				Version B									
score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICÍA benchmarks	score	freq.	theta	SE(th)	Benchmark (±1.5 units)	sdt	INICÍA benchmarks
0	0	-5,55	1,777				0	0	-5,728	1,779			
1	0	-4,41	0,915				1	0	-4,587	0,917			
2	0	-3,859	0,699				2	0	-4,034	0,701			
3	0	-3,482	0,592				3	0	-3,655	0,595			
4	0	-3,192	0,527				4	0	-3,362	0,529			
5	0	-2,952	0,481				5	0	-3,12	0,484			
6	0	-2,746	0,448				6	0	-2,911	0,451			
7	0	-2,565	0,422				7	0	-2,727	0,426			
8	0	-2,401	0,401				8	0	-2,56	0,405			
9	0	-2,251	0,385				9	0	-2,407	0,389			
10	0	-2,113	0,371				10	0	-2,265	0,375			
11	0	-1,983	0,359				11	0	-2,132	0,364			
12	0	-1,86	0,349				12	0	-2,007	0,354			
13	0	-1,744	0,341				13	0	-1,887	0,345			

14	0	-	1,633	0,333		14	0	-	1,772	0,338			
15	1	-	1,526	0,327	Exceptionally low		15	1	-	1,662	0,332		
16	0	-	1,423	0,322		16	1	-	1,555	0,327	Exceptionally low		
17	0	-	1,323	0,317		17	0	-	1,452	0,322			
18	1	-	1,226	0,313		18	0	-	1,351	0,319			
19	1	-	1,13	0,31		19	2	-	1,252	0,315			
20	3	-	1,037	0,307		20	4	-	1,155	0,313			
21	1	-	0,945	0,305		21	4	-	1,06	0,31			
22	5	-	0,854	0,303		22	10	-	0,965	0,309			
23	7	-	0,764	0,302		23	9	-	0,872	0,307			
24	6	-	0,675	0,301		24	9	-	0,779	0,307			
25	9	-	0,586	0,3		25	9	-	0,687	0,306			
26	16	-	0,497	0,301		26	11	-	0,594	0,306			
27	18	-	0,408	0,301		27	9	-	0,502	0,307			
28	15	-	0,319	0,302		28	12	-	0,409	0,307			
29	18	-	0,229	0,303		29	16	-	0,316	0,309			
30	19	-		0,305		insufficient	30	21	-	0,311			insufficient

		0,138						0,221			
31	29	-	0,308		sufficient	31	21	-	0,313		sufficient
32	19	0,048	0,311	Mediocre		32	20	-	0,316		
								0,029			
33	33	0,144	0,315			33	19	0,07	0,319	Mediocre	
34	32	0,243	0,319			34	20	0,172	0,323		
35	28	0,344	0,324			35	18	0,276	0,328		
36	16	0,448	0,33			36	30	0,384	0,334		
37	23	0,557	0,337			37	13	0,495	0,341		
38	17	0,671	0,345			38	18	0,611	0,349		
39	15	0,79	0,355			39	24	0,733	0,358		
Continuing...											
40	13	0,917	0,367		sufficient	40	23	0,862	0,369		sufficient
41	13	1,053	0,381		exceptional	41	17	0,999	0,383		exceptional
42	14	1,199	0,397			42	7	1,147	0,399		
43	4	1,359	0,418			43	4	1,308	0,419		
44	4	1,537	0,444	Exceptionally low		44	11	1,487	0,444		
45	2	1,739	0,477			45	5	1,689	0,477	Exceptionally low	
46	1	1,975	0,523			46	2	1,924	0,522		
47	0	2,261	0,589			47	0	2,21	0,588		
48	0	2,634	0,696			48	0	2,58	0,694		
49	1	3,182	0,912			49	0	3,124	0,91		
50	0	4,318	1,773			50	0	4,256	1,769		