"PROGRAMAS COMPUTACIONALES PARA EVALUACION DE TESTS"

BENEDICTO LOPEZ CATALAN

Universidad de Santiago de Chile
Depto. Matematica y Ciencia de
La Computación.

Colaboradores:
NANCY PEREZ J.
JUDITH AHUMADA G.

VIII ENCUENTRO NACIONAL DE INVESTIGADORES DE EDUCACION.

SANTIAGO - CHILE, SEPTIEMBRE 1985.

I N D I C E

PROGRAMAS COMPUTACIONALES PARA EVALUACION DE TESTS.

PROGRAMAS COMPUTACIONALES PARA EVALUACION DE TESTS

RESUMEN

El objetivo del presente trabajo es proporcionar una herramienta a los docentes a nivel de aula a los efectos de realizar análisis de test en medición educacional. Para ello se desarrolló en un microcomputador (*) un paquete computacional, que permite corregir y obtener indicadores asociados al instrumentos de medición.

El docente podrá así optimizar el tiempo destinado a la corrección del test.

Por otra parte los indicadores le darán pauta en torno a la calidad del instrumento empleado.

En forma muy breve en el tema que se presenta aquí, se desarrollarán las expresiones de los indicadores y entregará la metodología de uso de los programas generados.

(*) se usó microcomputador Apple II

I INTRODUCCION

Uno de los roles que debe asumir el educador es el de evaluador.

Enfrentada en su totalidad la tarea de evaluar, resulta amplia y compleja.

En nivel amplio el educador puede estar ligado a funciones que le exigen evaluar el proceso Educacional en un Ambito general, por ejemplo, evaluar la estructura curricular para un sistema de un país, o los resultados que produce el traspaso de colegios a a las municipalidades.

Por otra parte a nivel del aula el docente debe evaluar el proceso de instrucción para unidades de enseñanza y objetivos educacionales.

Otro aspecto que reviste la evaluación por parte del docente a nivel del aula es el de evaluar el rendimiento escolar.

Este tipo de evaluación dice relación con la medición educacional. Entrar en el proceso de medición del conocimiento que los estudiantes tienen en torno a una materia, implica haber evalua - do previamente objetivos que se desea lograr y los contenidos que se desea entregar.

Una de las fases de compromiso que tiene el educador ra

dica en la construcción de instrumentos que serán usados para realizar la medición. Estos instrumentos deben tener propiedades que garanticen su calidad.

El presente trabajo apunta en la dirección de conseguir una colección de indicadores que son expresiones cuantificables asociadas a la evaluación del instrumento de medición empleado.

Uno de los tipos de instrumentos de medición que suelen ser empleados por los docentes, es el de los tests objetivos.

Un criterio estadístico-matemático permite realizar análisis de las propiedades deseables en un test objetivo. Con ayuda de esta herramienta es posible también dar forma a los indicadores que son las expresiones de cuantificación.

Los métodos habituales de escritorio no facilitan la tarea de efectuar dicho análisis.

El avance de los dispositivos de Computación permite, generar un sistema de análisis de test, el cual da velocidad, seguridad y eficiencia para analizar instrumentos de medición de tipo objetivo.

En el desarrollo de este tema daremos una breve revisión a las propiedades deseables en un test, pasando luego a bosquejar cómo operar el archivo de programas preparados.

II CARACTERISTICAS PARA JUZGAR LA CALIDAD DE UN TEST

En forma muy breve pasaremos ahora a describir las características deseables en un test.

No entraremos a analizar los pasos que se debe dar para construir un test, dado que nuestro tema se refiere a los aspectos de cuantificación de expresiones que son indicadores de calidad.

Un test objetivo debe poseer las siguientes características : validez, confiabilidad, pertinencia, objetividad, equilibrio, especificidad, dificultad, discriminación, homogeneidad, practicidad y velocidad.

1.- Validez

La validez de un test está ligada al grado de precisión con que los items miden los rasgos que se pretende medir.

Los especialistas han planteado tres tipos de validez.

a) Validez predictiva

Este tipo de validez se refiere a la eficacia que tiene un test en la predicción de la conducta de un individuo en en situaciones específicas .

b) Validez concurrente

La validez concurrente es una característica empírica, desde el momento que los resultados del test deben ser comparados con un criterio externo, constituido por otra medición del mismo atributo, y que es realizada aproximadamente en el mismo tiempo. Suele utilizarse en situaciones de diagnóstico.

c) Validez de contenido

Este tipo de validez requiere identificar los objetivos específicos del curso o unidad en cuestión y examinar si el test los pone de manifiesto. Para el profesor a nivel de aula, éste será el tipo de validez más importante.

d) Validez de constructo

Un atributo o cualidad sicológica que suponemos que explica algún aspecto del comportamiento, es lo que llamamos un constructo. Ejemplos comunes de constructo son la inteligencia, actitud científica, comprensión de lectura.

La validez de constructo está ligada a la extensión y profundidad con que el test puede explicar los rasgos implicados por el constructo en cuestión.

2.- CONFIABILIDAD

La confiabilidad de un test dice relación con la consistencia que el test tiene como instrumento de medición. Si para un individuo se vuelve a medir el mismo rasgo en condiciones similares, los resultados deben ser los mismos.

3.- PERTINENCIA

Fijar una colección de criterios, precedidospor una declaración breve de objetivos es generar criterios de pertinencia en el test. Los criterios deben ser claros y puntuales a fin de limitar el test a los items del tipo deseado.

4.- OBJETIVIDAD

La objetividad de un item será también función de un criterio externo. Si los expertos en el tema coinciden en dar unanimemente la misma respuesta, decimos que el item tiene objetividad.

5.- EQUILIBRIO

Para juzgar el equilibrio de un test, debe especificarse de manera neta y clara el tipo de item apropiado para cada objetivo que se desee medir.

Una tabla de especificaciones facilitará en gran medida este trabajo.

6.- ESPECIFICIDAD

Un conocimiento específico requiere de una persona tener dominio del tema. Frente a un item específico los alumnos novicios no podrán enfrentarlo con éxito.

7.- DIFICULTAD

De un item de test podemos decir que es difícil, medianamente difícil o que es fácil. Esta es una propiedad ordinal que podrá ser estimada a priori y medida a posteriori para el grupo al cual se administró la pregunta.

8.- DISCRIMINACION

El grado en que un item sea capaz de diferenciar entre los sujetos que poseen un dominio más alto del tema respecto de aquellos que poseen poco dominio, genera el concepto de discriminación.

9.- HOMOGENEIDAD

Decimos que un test es homogéneo si con sus items se pretende medir el mismo rasgo sicológico en todos los individuos.

10.- PRACTICIDAD

Un buen test no puede estar exento de características que los tornen práctico. Factores tales como : tiempo de preparación, costo, administración y revisión deberán ser considera-

dos como elementos ligados a la practicidad.

11.- Velocidad

Esta propiedad puede también estimarse a priori y constatarse a posteriori. Un test se considera bien planeado si él es respondido por los alumnos en el plazo estipulado.

Se piensa que es recomendable un tiempo tal, que el 90 % de los sujetos pueda intentar responder la última pregunta del test.

De las características señaladas cinco son cuantificables esto es, poseen indicadores.

Ellas son : confiabilidad, validez (*), dificultad, discriminación y homogeneidad.

(*) Un coeficiente de validez tiene serios problemas de interpretación y por ello se prefirió no programarlo.

III BASES DE CUANTIFICACION

Matriz de puntajes de test

Supongamos un test de n items que es presentado a m sujetos para ser respondido.

Cada individuo en cada item tiene opción de responder acertadamente, no acertar en su respuesta, o no contestar.

Podemos visualizar una tabla de doble entrada o matriz de puntajes, para la cual convenimos en asignar 1 a cada linea de cruce de respuestas acertada y 0 en otro caso.

Simbolicemos la matriz de puntajes de la siguiente forma :

$$P = (x_{ij})_{m \times n}$$

$$x_{ij} = \begin{cases} 1 \text{ si el individuo i-ésimo, responde acertadamente el item j-ésimo} & i = 1,2 \ldots m \\ & j = 1,2 \ldots n \\ 0 \text{ si el individuo i-ésimo, responde no acertamente o no responde el item j-ésimo} \end{cases}$$

Para cada individuo el test aparece así como una colección de n pruebas de Bernouilli.

Representemos por $x_{i.}$ el total de respuestas acertadas por el individuo i, i= 1,2... m

$$x_{i.} = \sum_{j=1}^{n} x_{ij}$$

Suponiendo que se dan ciertos supuestos matemáticos podemos aceptar que $x_{i.}$ tiene distribución de tipo binomial.

Si el número m de individuos es grande podremos usar una aproximación normal para la distribución binomial de "Puntajes". No nos corresponde ahora entrar en mayores detalles desde el punto de vista matemático.

Consideremos también la expresión, correspondiente al total de respuestas dadas al item j j = 1,2,... n

Denotemos esta expresión por $x_{.j}$

entonces $x_{.j} = \sum_{i=1}^{m} x_{ij}$ j=1,2,...n

Desde el punto de vista frecuencial podemos entonces estimar la probabilidad de respuesta al item j por el grupo de m individuo sometidos al test.

Tendremos : Prob { respuesta acertada item j} $= \frac{x_{.j}}{m} = p_j$

j = 1,2,..m

Por otra parte p_j mide el número medio de respuestas al item j.

Con un planteamiento de experimento dicotómico, podemos caracterizar $q_j = 1 - p_j$

q_j = Prob { respuesta no acertada o no respuesta al item j }

La varianza del item j, s_j^2 será $p_j\ q_j$

Modelo clásico de puntuación

El puntaje que un individuo acumula en un test y que denotamos por x_i. puede ser considerado como constituídos por dos componentes; por una parte una expresión que corresponde a valor verdadero T_i y por otra el error que se comete al observar al individuo : e_i.

Si usamos este esquema clásico tendremos un modelo regresivo de la forma :

$t = T + e$ t : puntaje observado

T : puntaje verdadero

e : error de medición

Dado que este modelo es una regresión será válido plantear lo siguiente :

El error es una variable de media nula y varianza σ_ϵ^2 y no correlacionado con la variable T.

Para cada individuo se tendrá : $t_i = T_i + e_i$

$i = 1,2...m$

Test paralelos

Parte de nuestro estudio estará basado en la idea detests paralelelos.

Consideraremos para una colección de objetivos y contenidos de una unidad educacional el conjunto U, de todos los posibles items asociados a los objetivos y contenidos mencionados.

Si se desea construir un test para medir a un grupo de individuos sobre los objetivos planteados, este test X será una muestra aleatoria de tamaño n , extraída del conjunto U. $\bar{X}$ es una forma de test, pero no la única. El conjunto de forma de test equivalentes con X constituirá una clase de equivalencia, digamos $\mathfrak{X}$, a la que llamaremos el conjunto de test paralelos. Notemos que la relación ser paralelos entre test constituye una relación de equivalencia.

Sean X y Y dos formas paralelas de test. Un individuo sometido a cualquiera de las dos formas deberá tener la misma puntuación y con la misma variabilidad. Desde un punto de vista distribucional para un grupo de individuos los puntajes deberán tener el mismo comportamiento.

Necesitaremos las ideas anteriores para efectuar análisis de las medidas asociadas a las características de un test.

Confiabilidad

Supuesto que (para un individuo) se dispone de dos mediciones en torno a un rasgo.

$$t_{i_1} = T_i + e_{i_1} \qquad i = 1,2...m$$

$$t_{i_2} = T_i + e_{i_2}$$

La confiabilidad entre las medidas t_1 y t_2 quedará expresada como :

$$r_{t_1 t_2} = \frac{s_T^2}{s_t^2}$$

Si consideramos el test completo tendremos

$$r_{tt} = \frac{s_t^2 - s_e^2}{s_t^2} = 1 - \frac{s_e^2}{s_t^2}$$

Para aproximarnos a la estimación de la confiabilidad, hagamos uso de la fórmula de Profecía de Spearman Brown, la que nos lleva a :

$$r_{tt_n} = \frac{n\, r_{tt}}{1 + (n-1)\, r_{tt}}$$

Métodos para estimar la confiabilidad

1. Coeficiente α de Crombach (1951)

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{1}^{n} s_j^2}{s_t^2}\right)$$

s_t^2 : Varianza de puntajes de test

s_j^2 : Varianza de puntajes de item j. j=1,2...n

2. Método común de división por mitades

Si en un test obtenemos dos subtest, X el de los items impares y Y el de los items pares r_{xy} representará la correlación entre ambos subtests y el coeficiente de confiabilidad estará dado por :

$$r_{tt} = \frac{2\, r_{xy}}{1 + r_{xy}}$$

X: representa items impares

Y: representa items pares

3. Método de Rulon-Guttman

En este caso se hace división de item impares (X), items pares (Y) formando dos subtest. Tomando las diferencias de puntajes (d). se determina el coeficiente de confiabilidad como :

$$r_{tt} = 1 - \frac{s_d^2}{s_t^2}$$

s_d^2 : es la varianza de diferencias de puntajes en los subtest.

4. Método de Kuder y Richardson

$$r_{tt} = \frac{n}{n-1}\left\{\frac{s_t^2 - \sum_1^n p_j q_j}{s_t^2}\right\} \quad [KR - 20]$$

La expresión $\sum_1^n p_j q_j$ puede sustituirse por $n\,\bar{p}\bar{q}$ obteniendo.

$$r_{tt} = \frac{n}{n-1}\left\{\frac{s_t^2 - n\,\bar{p}\,\bar{q}}{s_t^2}\right\} \quad [KR - 21]$$

Esta última expresión es equivalente a :

$$r_{tt} = \frac{n}{n-1}\left\{1 - \frac{\bar{x}(n - \bar{x})}{n\,s_t^2}\right\}$$

5. Método de Hoyt

Hoyt parte de la siguiente relación

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad \begin{matrix} i = 1,2...m \\ j = 1,2...n \end{matrix}$$

x_{ij} : puntaje del individuo i en item j .

μ : media poblacional

α_i : efecto persona

β_j : efecto item

θ_{ij} : error de medición

Los suspuestos del modelo son los supuestos habituales en el análisis de Varianza, llegando a la siguiente tabla :

A N O V A

FUENTE	SS	D.F.	MS	ESPERANZA DE M.S.
Persona	SS_p	m-1	$\frac{SS_p}{m-1}$	$\sigma^2 + \frac{n}{m-1} \sum_{i=1}^{m} t^2$
Item	SS_I	n-1	$\frac{SS_p}{n-1}$	$\sigma^2 + \frac{m}{n-1} \sum_{j=1}^{m} p_j^2$
Error	SS_e	(m-1) (n-1)	$\frac{SS_e}{(m-1)(n-1)}$	σ^2
Total	SS_T	mn-1		

La estimación del coeficiente de confiabilidad es :

$$r_{tt} = 1 - \frac{\frac{SS_e}{(m-1)\ (n-1)}}{\frac{SS_p}{m-1}}$$

Discriminación del test (δ)

$$\delta = \frac{(n+1)\quad (m^2 - \sum_{i=1}^{n} n_i^2)}{n \quad m^2} -$$

n : número de items del test

m : número de individuos

n_i : frecuencia absoluta de puntajes de test

Indice de dificultad (D)

Para el item j denotaremos por d_j el indice de dificultad y será el cuociente entre el número de alumnos que responden acertadamente y el total de alumnos.

$$d_j = \frac{x.j}{m} \cdot 100 \qquad j = 1,2...m$$

$$D = \frac{1}{n} \sum_{j=1}^{n} d_j$$

Indice de homogeneidad (Loevinger) (H_t)

Para un test de n items al cual son sometidos m individuos Loevinger estructura el siguiente indice para medir su grado de homogeneidad.

$$H_t = \frac{m\left(\sum_{i=1}^{m} x_{i\bullet}^2 - \sum_{i=1}^{m} x_{i\bullet}\right) + \sum_{j=1}^{n} x_{\bullet j}^2 - \left(\sum_{i=1}^{m} x_{i\bullet}\right)^2}{2m\left(\sum_{j=1}^{n} j\, x_{\bullet j} - \sum_{i=1}^{m} x_{i\bullet}\right) + \sum_{j=1}^{n} x_{\bullet j}^2 - \left(\sum_{i=1}^{m} x_{i\bullet}\right)^2}$$

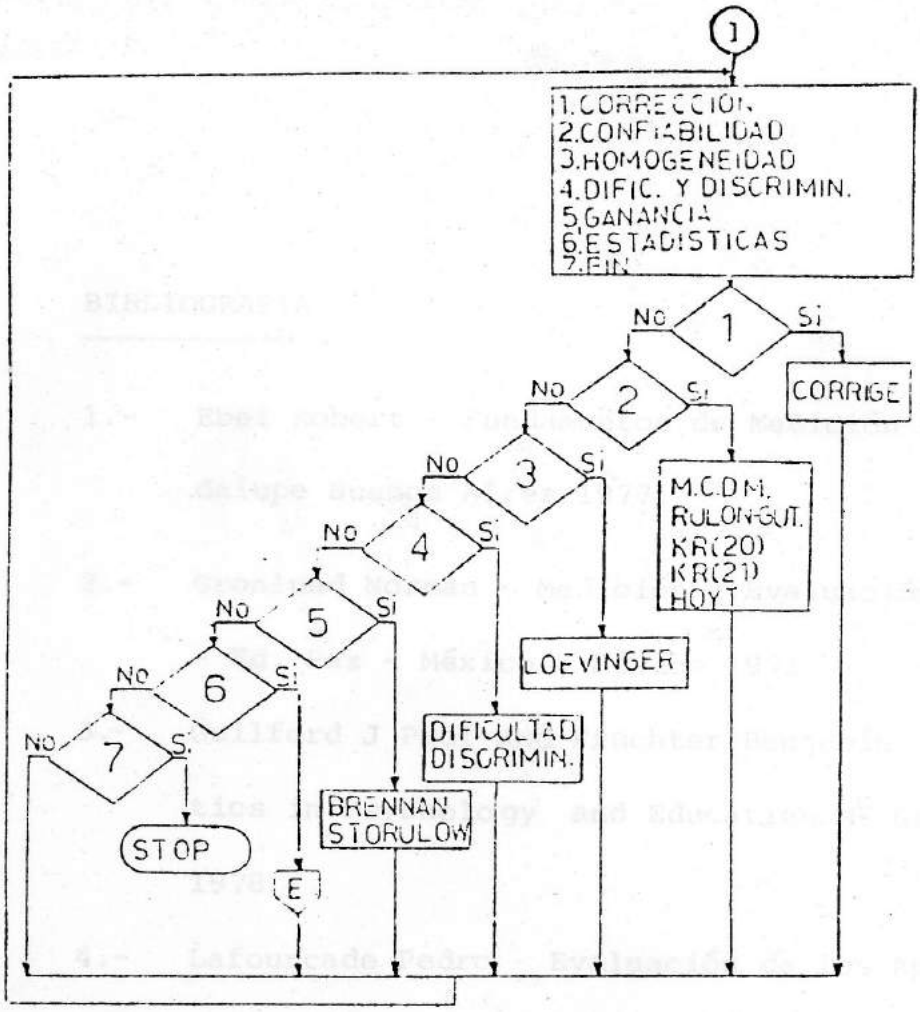
Notemos que este coeficiente es independiente de validez y confiabilidad.

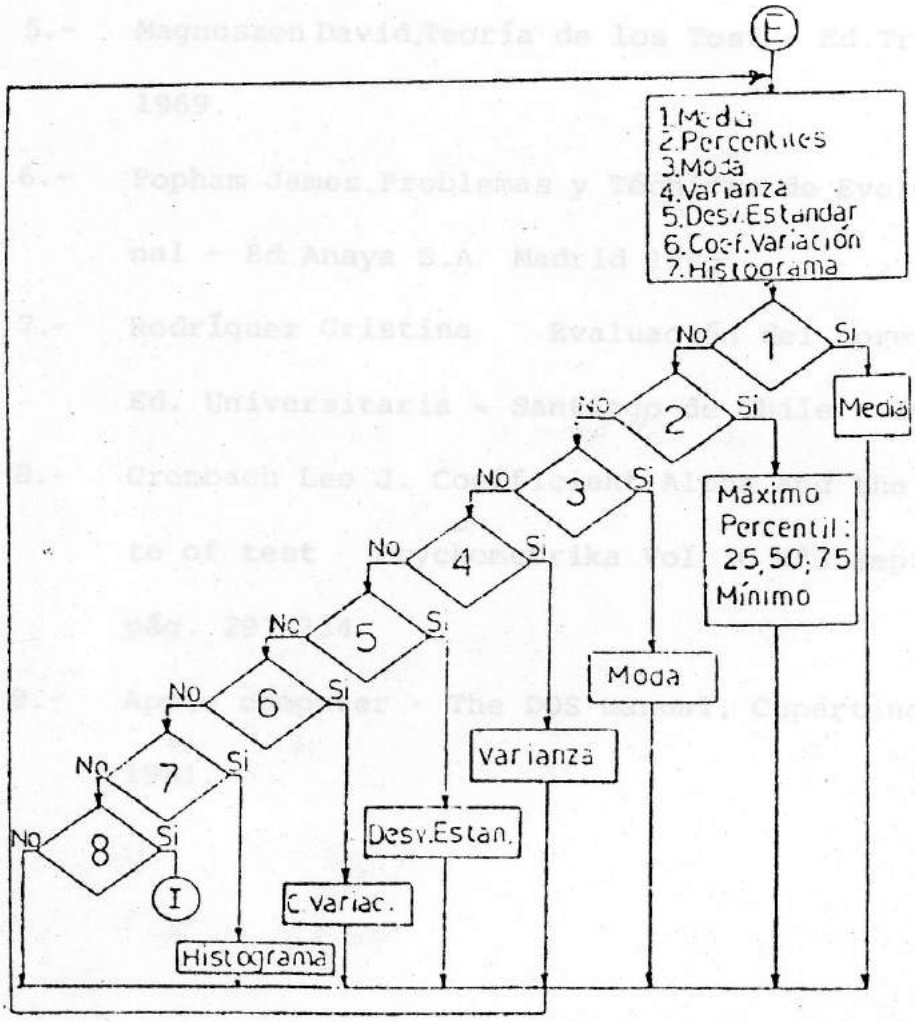
IV ARCHIVO DE PROGRAMAS COMPUTACIONALES

Los índices referenciados antes, las estadísticas básicas de puntajes de test y un programa de corrección de test fueron programados y con ellos se confeccionó un archivo factible de ser usado por los docentes.

Los programas fueron confeccionados en lenguaje BASIC y se usó un microcomputador.

El siguiente diagrama de flujo proporciona, trayectorias de uso de los programas.


1
1.CORRECCION
2.CONFIABILIDAD
3.HOMOGENEIDAD
4.DIFIC. Y DISCRIMIN.
5.GANANCIA
6.ESTADISTICAS
7.FIN
1
No
Si
CORRIGE
2
No
Si
M.C.D.M.
RULON-GUT.
KR(20)
KR(21)
HOYT
3
No
Si
LOEVINGER
4
No
Si
DIFICULTAD
DISCRIMIN.
5
No
Si
BRENNAN
STORULOW
6
No
Si
E
7
No
Si
STOP


E
1.Media
2.Percentiles
3.Moda
4.Varianza
5.Desv.Estandar
6.Coef.Variación
7.Histograma
1
No
Si
Media
2
No
Si
Máximo
Percentil:
25,50,75
Mínimo
3
No
Si
Moda
4
No
Si
Varianza
5
No
Si
Desv.Estan.
6
No
Si
C.Variac.
7
No
Si
Histograma
8
No
Si
I

BIBLIOGRAFIA

1.- Ebel Robert - Fundamentos de Medición Educacional-Ed. Guadalupe Buenos Aires 1977.

2.- Gronlund Norman - Medición y Evaluación de la enseñanza - Ed. Pax - México , México 1973.

3.- Guilford J Paul and Fruchter Benjamín- Fundamental Statistics in Psychology and Education. Mc Graw Hill, New York 1978.

4.- Lafourcade Pedro - Evaluación de los Aprendizajes. Ed. Kapelusz, Buenos Aires, 1973.

5.- Magnusson David Teoría de los Test - Ed.Trillas México - 1969.

6.- Popham James Problemas y Técnicas de Evaluación Educacional - Ed Anaya S.A Madrid 1980.

7.- Rodríguez Cristina Evaluación del aprendizaje escolar Ed. Universitaria - Santiago de Chile - 1978.

8.- Crombach Lee J. Coefficient Alpha and the internal Structe of test - Psychometrika Vol 16 N°3 September - 1951 pág. 297-334

9.- Apple computer - The DOS manual, Cupertino, California - 1981.