

## UN TEST ADAPTATIVO INFORMATIZADO DE ANÁLISIS LÓGICO BASADO EN LA GENERACIÓN AUTOMÁTICA DE ÍTEMS

Javier Revuelta y Vicente Ponsoda  
Universidad Autónoma de Madrid

En este artículo se describen los primeros pasos para transformar un test de lápiz y papel en un test adaptativo basado en la generación automática de ítems. El artículo tiene tres secciones principales: en la primera se describe la estimación de un modelo psicométrico componencial a partir del test de lápiz y papel, la segunda trata sobre la creación de un banco de ítems aplicando mecanismos generativos, y en la tercera se compara la precisión del test de lápiz y papel con la de un test adaptativo creado a partir de los ítems generados automáticamente.

*A logical analysis computerized adaptive test based on automatic item generation.*  
The initial efforts to transform a paper and pencil test into a computerized test based on automatic item generation are described in this paper. It has three main sections: Section one describes a psychometric model and its use to predict item difficulties. Section two describes a mechanism to generate new items with a given difficulty. Finally, section three compares the two tests (the based on automatic item generation and the paper and pencil test) with regard to ability estimation precision.

La influencia creciente de la Psicología cognitiva en la construcción de tests ha provocado un mayor interés en las operaciones mentales y procesos llevados a cabo por los sujetos que responden los tests, con el objetivo de lograr una mejor comprensión de los constructos que miden dichos tests (Embretson, 1994). Los análisis cognitivos pueden utilizarse para crear bancos de ítems de características psicométricas conocidas (Snow y Lohman, 1993), sin necesidad de calibrar los ítems en una muestra de sujetos reales,

en lo que se conoce como Generación Automática de Ítems (GAI).

A su vez estos bancos pueden utilizarse para crear Tests Adaptativos (TAIs) (Wainer, 1990) basados en la GAI. Algunos ejemplos de estas técnicas pueden verse en Collis, Tapsfield, Irvine, Dann y Wright (1995), Bejar (1993) y Hornke y Habon (1986).

En este artículo se describen los esfuerzos iniciales para transformar un test disponible comercialmente, el DA5, en un TAI basado en la GAI. El objetivo es servir de ejemplo sobre los pasos necesarios y mostrar algunas de las dificultades, teóricas y prácticas, que surgen al aplicar la GAI, de modo que resulte útil para otros investigadores que deseen aplicar las mismas técnicas. En primer lugar, se describe el test y un

---

Correspondencia: Vicente Ponsoda  
Facultad de Psicología  
Universidad Autónoma de Madrid  
28049 Madrid (Spain)  
E-mail: Vicente.Ponsoda@uam.es

modelo cognitivo de cómo se resuelven los ítems. A continuación, se presenta un modelo psicométrico para predecir la dificultad de los ítems. En tercer lugar, se describe un procedimiento para generar todos los ítems posibles para este test. Finalmente, se compara la precisión de un test adaptativo creado a partir del banco de ítems generados automáticamente con la del test de lápiz y papel.

Descripción del test DA5

El test DA5 ha sido desarrollado por la compañía SHL (1996) para medir análisis lógico mediante la habilidad de seguir instrucciones complejas. El test contiene 50 ítems ordenados por dificultad y tiene que ser completado en un tiempo de 20 minutos. En la figura 1 puede verse un ítem similar a los del test DA5. La columna izquierda contiene diversas figuras, cada una dentro de un cuadrado. A la derecha de cada figura aparece una instrucción simbólica dentro de un círculo. Hay una instrucción adicional dentro de un rombo al final de la columna de instrucciones simbólicas. La figura 1 muestra un ítem de cuatro filas. El test DA5 consta de ítems de 2, 3 y 4 filas.

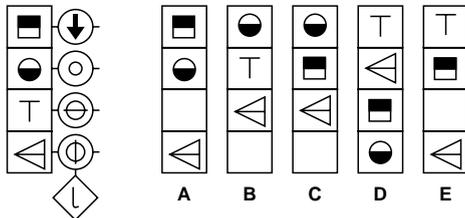


Figura 1. Ejemplo de ítem del test DA5.

En el test aparecen 10 instrucciones diferentes, que indican cómo ha de transformarse mentalmente la figura adyacente. Unas indican que la figura ha de ser rotada; otras, que se ha de hacer un intercambio en

tre figuras adyacentes. La instrucción adicional dentro del rombo indica cómo debe reordenarse mentalmente toda la columna de figuras. El sujeto dispone de una hoja en la que se describe el significado de cada instrucción y cómo ha de proceder. Ha de aplicar una a una todas las instrucciones, comenzando por la de arriba, y determinar cómo quedaría la columna de figuras después de estas transformaciones. Finalmente, ha de escoger su respuesta entre las columnas marcadas con las letras de la 'A' a la 'E'.

Un modelo cognitivo elemental

Se propone un modelo cognitivo muy sencillo de la forma en que los sujetos resuelven la tarea. El modelo se basa en dos supuestos: 1) Las instrucciones del ítem se aplican de forma secuencial, empezando por la fila superior. 2) El procesamiento es exhaustivo; es decir, se aplican todas las instrucciones. A partir de estos dos supuestos, el modelo consiste en la repetición de los siguientes pasos hasta agotar todas las instrucciones: 1) codificación de una figura y una instrucción, y 2) aplicación de la instrucción sobre la figura. Finalmente el sujeto debe elegir una de las alternativas de respuesta.

Para determinar los componentes de dificultad se añade un supuesto adicional. De los distintos componentes del modelo (codificación, aplicación de instrucciones y elección de la respuesta), se asume que la codificación y elección de la respuesta no son causa de diferencias individuales en rendimiento en el ítem. Estas diferencias se deben exclusivamente al proceso de aplicación de las instrucciones. Por lo tanto, se hipotetiza que la dificultad de los ítems de este test se puede descomponer en la dificultad que implica la aplicación de cada una de las instrucciones.

Un modelo psicométrico para predecir la dificultad

Además de identificar los componentes de dificultad es necesario relacionarlos con la dificultad del ítem, esto puede hacerse mediante la utilización de un modelo exponencial de la Teoría de Respuesta al Ítem (TRI). Es decir, uno en el que la dificultad no se asigne a los ítems individuales sino a sus componentes.

Uno de los modelos exponenciales más utilizados es el modelo logístico lineal de rasgo latente (Linear Logistic Latent Trait Model, LLTM), el cual se basa en el modelo de Rasch:

$$(1) \quad P(x = 1 | \theta, b_i) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Donde el parámetro de dificultad del ítem ( $b_i$ ) se descompone de forma lineal en los parámetros de dificultad de cada uno de los componentes (Fischer y Molenaar, 1995):

$$(2) \quad b_i = \sum_{x=1}^m p_{xi} q_x + d$$

' $b_i$ ' es la dificultad del ítem 'i', 'm' es el número de componentes, ' $p_{xi}$ ' frecuencia con que aparece el componente 'x' en el ítem 'i', ' $q_x$ ' es la complejidad del componente 'x' y 'd' es una constante de normalización. También se empleó el modelo logístico de 3 parámetros (3pl) debido a que al incorporar una asíntota inferior puede ajustar mejor a datos procedentes de ítems de respuesta múltiple (Hambleton y Swaminathan, 1985), que se pueden acertar si se responde al azar:

$$(3) \quad P(x = 1 | \theta, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

Donde ' $a_i$ ' es la capacidad discriminativa del ítem 'i' y ' $c_i$ ' es el parámetro de pseudoazar. El modelo LLTM se empleó para comprobar la hipótesis de que la dificultad de los ítems del test DA5 puede descomponerse en la de las instrucciones que intervienen. Por lo tanto, 'm' es igual al número de posibles instrucciones (10) y 'p' es la frecuencia con que aparece cada una de estas instrucciones en cada ítem.

Método

*Sujetos y material*

Se empleó una muestra de 621 sujetos que habían respondido al test DA5, proporcionada por SHL-España. Los sujetos respondieron al test durante varios procesos de selección de personal.

*Análisis de datos*

En primer lugar se muestran algunos estadísticos de la Teoría Clásica de Tests. A continuación, se utiliza el procedimiento propuesto por Bejar (1980) para comprobar la unidimensionalidad del test DA5, consistente en estimar los parámetros del modelo por separado en el test total y en distintos subconjuntos de ítems: por ejemplo los ítems pares y los impares. Si el test es unidimensional, los parámetros deben ser similares en las dos calibraciones.

A partir de la matriz de respuestas se determinó un subconjunto de sujetos e ítems en el que se ajustase razonablemente bien el modelo de Rasch, con el objeto de estimar el modelo LLTM en esta muestra reducida. La frecuencia de aparición de cada una de las 10 instrucciones se utilizó como predictor de la dificultad del ítem.

*Resultados*

La proporción de aciertos de los ítems oscila entre 0.04 y 0.99, con media 0.64. Por

tanto, los ítems no resultan excesivamente difíciles, sin embargo existe una fuerte relación entre la proporción de aciertos y la posición en el test ( $r=-0.84$ ,  $p<0.000$ ). La correlación biserial ítem-total oscila entre 0.15 y 0.88, con media 0.62. Este último dato indica que puede ser necesario estimar un modelo de TRI con un parámetro de discriminación diferente para cada ítem (Lord, 1980). Respecto a las características del test completo, la puntuación media es 32, y la desviación típica 8, lo cual confirma que el test tiene una dificultad media para el grupo de sujetos. La consistencia interna es elevada, obteniéndose un  $\alpha$  igual a 0.90.

La correlación entre los parámetros de dificultad estimados en el test total y en la mitad correspondiente resultó ser 0.99, siendo además similar la magnitud de los mismos. Por tanto puede concluirse que, según el procedimiento de Béjar, el test tiene una unidimensionalidad aceptable.

Se ajustó el modelo de Rasch y el 3pl a toda la muestra y a todos los ítems del test mediante los programas RASCAL y ASCAL (ASC, 1994). El ajuste al modelo de Rasch es moderado, con valores de  $\chi^2$  entre 16 y 70 (con 19 gl.). Por el contrario, para 8 de los ítems dicho estadístico toma valores claramente superiores al resto. La inclusión de los parámetros de pseudoazar y discriminación en el 3pl mejora sustancialmente el ajuste para todos los ítems, aunque los ítems que peor ajustan al modelo de Rasch también son los que peor ajustan al 3pl. No se encontró una relación clara entre el ajuste del modelo de Rasch para cada ítem y los estadísticos clásicos, ni tampoco con los valores estimados para los parámetros 'a' y 'c'. En cambio, en el grupo de ítems en los que el ajuste es claramente peor al del resto, se encuentra que la proporción de aciertos es alta y el parámetro 'b' (tanto de Rasch como del 3pl) es bajo, en torno a -2; en estos ítems las proporciones observadas de aciertos no siempre aumentan a medida lo que hace la habilidad.

Debido al bajo ajuste del modelo de Rasch se intentó encontrar un subconjunto de la muestra total de sujetos en los que este modelo produjese un mejor ajuste, con la intención de estimar los parámetros del LLTM en esta muestra reducida. Para ello se utilizó el índice  $z^2$  de ajuste de la persona 's' al ítem 'i', propuesto por Wright y Stone (1979, p. 69 y ss.):

$$(4) \quad z_{si}^2 = \exp\left[(2r_{si} - 1) (b_i - \theta_s)\right]$$

Donde  $r = 1$  si la respuesta es correcta, y 0 si es incorrecta. A cada combinación persona x ítem le corresponde valor de  $z^2$ , estos valores pueden sumarse a lo largo de las personas o de los ítems para obtener un indicador del ajuste de cada persona e ítem al modelo, distribuido según  $\chi^2$  con T-1 grados de libertad (donde T es el número de términos que intervienen en la suma). Este índice puede utilizarse para obtener una muestra reducida de personas, en las que ajuste el modelo de Rasch, y estimar el ella el LLTM (Hornke y Habon, 1986). Se tomó la decisión de eliminar aquellos sujetos de la muestra cuyo valor  $z_s^2$  fuera significativo al 1% ( $_{0.99}\chi^2_{49} \approx 76$ ). De esta forma se obtuvo una muestra de 279 personas, en la cual 7 de los ítems no tenían varianza. Se prefirió incrementar el ajuste mediante reducciones de la muestra, y evitar en lo posible la reducción en el número de ítems debido a que el número de componentes es alto (10) en relación al número de ítems. Para obtener los parámetros del modelo LLTM se utilizó el programa LPCM-WIN (Fischer y Ponocny-Seliger, 1997) en la muestra final de 43 ítems y 279 personas. La tabla 1 muestra los principales resultados relativos a la bondad del ajuste. En lo sucesivo, al referirnos al test DA5 lo haremos al test DA5 de solo 43 ítems.

*Tabla 1*  
Parámetros estimados y bondad de ajuste de los modelos 3pl, Rasch, y Rasch en la muestra reducida

Variable	Media	Std.	Mín.	Máx.	N.
a 3pl	1.12	0.64	0.40	2.38	50
b 3pl	-0.81	1.48	-3.00	2.09	50
c 3pl	0.20	0.04	0.11	0.31	50
$\chi^2$ 3pl (17 gl.)	29.17	31.89	2.85	226.09	50
b Rasch	0.00	2.36	-4.03	5.48	50
$\chi^2$ Rasch (19 gl.)	187.67	753.31	16.31	5267.86	50
b Rasch	0.00	1.00	-2.84	3.90	43
$\chi^2$ Rasch (19 gl.)	17.27	8.11	3.44	43.89	43

La correlación entre los parámetros 'b' del modelo de Rasch y los predichos por el LLTM resultó ser 0.684. Es decir, considerando las posibles instrucciones que pueden aparecer en cada ítem es posible explicar un 47% de la varianza en dificultad de los ítems. Las instrucciones que tienen un mayor efecto en la dificultad son la 9 y la 10, que implican la reordenación de un vector de 4 figuras (ver tabla 2).

*Tabla 2*  
Parámetros estimados del LLTM. 'q' es el parámetro de dificultad de cada componente,  $Se_q$  el error típico de estimación. Z es el estadístico de contraste y P(Z) su probabilidad asociada

Variable	q	$Se_q$	Z	P(Z)
p1	0.0334	0.042	6.317	0.000..
p2	0.0929	0.040	18.332	0.000..
p3	0.0126	0.081	1.244	0.1075
p4	0.1002	0.066	12.036	0.000..
p5	0.1153	0.037	24.550	0.000..
p6	0.0803	0.066	9.693	0.000..
p7	0.1355	0.084	12.828	0.000..
p8	0.1372	0.054	20.181	0.000..
p9	0.4122	0.100	32.711	0.000..
p10	0.4903	0.101	38.724	0.000..
d	-0.0625			

Por último, debe destacarse que en entrevistas realizadas con 13 sujetos que realiza-

ron el test, 9 de ellos informan que aplican todas las instrucciones del ítem (cómo asume el modelo), sin embargo, los otros 4 aplican solamente el mínimo número de instrucciones necesario para alcanzar la solución correcta. Además, se encuentra un efecto de interacción entre el tipo de estrategia utilizada y la longitud del ítem, aplicándose todas las instrucciones en los ítems cortos pero no en los de cuatro filas. Estos resultados indican la necesidad de una mayor elaboración del modelo cognitivo y la consideración de estrategias alternativas de respuesta. A pesar de los problemas mencionados, en el resto del artículo se utilizará la ecuación (2) para predecir el parámetro de dificultad del modelo de Rasch de un banco completo de ítems, creado según la GAI. El propósito es ilustrar las ventajas que pueden obtenerse si se aplica esta tecnología en un TAI.

### Generación de Ítems

Una posibilidad para desarrollar un generador de ítems es utilizar la Teoría de Automatas (Brookshear, 1989). En el test DA5 un ítem puede describirse formalmente como una matriz de tamaño  $R \times C$ , donde R es el número de filas del ítem (5) y C el de columnas (7). Esta matriz contiene distintos símbolos que representan a cada una de las figuras e instrucciones que aparecen en el ítem. Además es necesario especificar un conjunto de reglas para colocar los símbolos en la matriz. El número de reglas es demasiado grande para describirlas en este artículo, aunque pueden agruparse dependiendo de la función que realizan. Para generar un ítem se aplican secuencialmente los distintos grupos de reglas:

- 1) Creación de la columna inicial de figuras. Estas figuras se seleccionan aleatoriamente de entre las 36 posibles.
- 2) Creación de una columna admisible de instrucciones

(por ejemplo, no puede colocarse una instrucción que indique «rotar la figura de la siguiente fila» en la última fila). 3) Reglas para generar la alternativa correcta. Estas reglas aplican las instrucciones del ítem sobre las figuras iniciales, la columna de figuras resultante es la alternativa correcta. 4) Reglas para crear las alternativas incorrectas. Las alternativas incorrectas empleadas en el DA5 comparten las mismas figuras que la correcta, aunque su colocación o las rotaciones de algunas figuras pueden variar. Para generar alternativas incorrectas que también posean estas características, las reglas funcionan del siguiente modo: en primer lugar se modifica aleatoriamente alguna de las instrucciones del ítem, por ejemplo se cambian los grados de alguna rotación o se anulan instrucciones de cambio de orden de figuras. El vector modificado de instrucciones se aplica sobre las figuras originales, y se comprueba que las figuras resultantes de la aplicación son diferentes de la alternativa correcta.

El número total de ítems que puede crearse con estas reglas es 4242. Si además de las instrucciones se modificasen sistemáticamente las figuras del ítem, el número de combinaciones sería considerablemente mayor. A continuación se compara la eficiencia de un TAI basado en la GAI, según la gramática descrita, con el test DA5 original. El propósito de este estudio es simplemente ilustrativo sobre las ventajas de la GAI (Fischer y Pendl, 1980), debido a que todavía no se cuenta con un modelo psicométrico para predecir la dificultad con una precisión suficiente.

Precisión del test generado automáticamente

#### Método

##### *Procedimiento*

Se crearon los 4242 ítems posibles y su parámetro de dificultad del modelo de

Rasch se predijo según la ecuación (2). Se compararon los parámetros de los ítems y las funciones de información del test DA5 (de 43 ítems), del banco generado, de un TAI de 43 ítems creado a partir del banco generado y del test DA5.

#### Resultados

Según indican los estadísticos descriptivos, el banco de ítems generado cubre un rango de dificultades mucho menor que el test DA5. Los valores extremos de la dificultad son -2.84 y 3.9, en el test DA5, y -0.03 y 0.95, en el banco generado automáticamente. Esto se debe a que el LLTM no ha resultado eficaz para predecir las dificultades extremas. Además, el promedio de dificultad es superior en el banco generado (0.5) que en el DA5 (0.0), lo que indica que en este último test no se ha utilizado una muestra aleatoria de todos los ítems posibles, sino que se ha dado preferencia a la inclusión de ítems fáciles.

La figura 2 muestra que si se divide la información por el número de ítems, cada ítem del DA5 es más informativo que los del banco generado para los niveles de habilidad menores que cero, mientras que los ítems generados son más informativos, en promedio, para los niveles mayores que cero. Este resultado se debe a las diferencias comentadas en la distribución del parámetro de dificultad en ambos bancos: El banco generado apenas tiene ítems con dificultades menores que cero por lo que resulta poco informativo para capacidades inferiores a cero. Por tanto, puede concluirse que aunque la generación de ítems puede ser útil para mejorar la eficacia del test; sin embargo, es necesario predecir adecuadamente la dificultad de los ítems con dificultades extremas antes de utilizarla con garantías (ver figura 2).

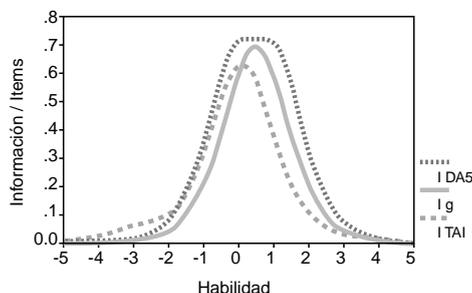


Figura 2. Información media por ítem en relación con el nivel de habilidad para el test DA5, banco de ítems generado y un test adaptativo creado a partir del banco generado.

### Conclusiones

Este artículo describe los pasos iniciales para transformar un test de lápiz y papel en un test adaptativo basado en la generación automática de ítems. Sin embargo, resulta necesario subsanar muchas carencias antes de poder contar con un test definitivo. En primer lugar, es necesario comprobar que los ítems generados tienen las propiedades psicométricas predichas por el modelo. En segundo lugar, la proporción explicada de varianza en dificultad de los ítems es alta pero no suficiente para aplicar la GAI, y es especialmente necesario mejorar la predicción de los ítems con dificultades extremas. Una forma de mejorar esta predicción puede ser añadir nuevos predictores que incluyan información no sólo de las instrucciones del ítem sino de las figuras que aparecen en todas las alternativas.

Un inconveniente del LLTM es la falta de una asíntota inferior, lo que puede producir una falta de ajuste cuando se emplean ítems de alternativas múltiples. Mediante el LLTM es posible predecir la dificultad de los ítems, sin embargo no se ha desarrollado un modelo exponencial que permita predecir la asíntota inferior y la capacidad discriminativa a partir del contenido del ítem. Otra dificultad de la GAI es la variación en

los parámetros de dificultad de los componentes, en función de la muestra de ítems empleada. Es posible que utilizando una muestra amplia de ítems se consigan errores bajos en la estimación; no obstante, esto no garantiza que no existan subconjuntos del banco total de ítems para los cuales los parámetros estimados de dificultad puedan ser muy diferentes. Esta puede ser una causa de un desajuste entre la dificultad real de los ítems que se administran a un sujeto concreto durante un test basado en la GAI y la dificultad predicha por el modelo psicométrico. Una segunda causa es la existencia de estrategias de resolución de los ítems diferentes a la considerada durante la estimación del modelo exponencial.

Hay todavía otra dificultad sobre la que conviene llamar la atención. Un paso previo a la aplicación del LLTM es comprobar que el modelo de Rasch ajusta a los datos. Por esa razón, se han eliminado del test DA5 sujetos e ítems para incrementar el ajuste. Sin embargo, cuando se genera el banco completo, se generan todos los ítems, tanto los que se ajustan al modelo, como los que no. Nos encontramos, entonces, con la paradoja de que hemos establecido a partir de unos pocos «buenos» ítems la correlación entre las dificultades reales y las predichas a partir de sus componentes, pero que los valores utilizados para obtener las dificultades predichas se aplican a *todos* los ítems; tanto a los que se ajustan al modelo, como a los que no. El asunto es especialmente importante si se repara en que en un TAI basado en la GAI se aplican ítems creados expresamente para cada sujeto, sobre los que no se posee información sobre su ajuste al modelo.

Los resultados de este estudio indican que la GAI puede resultar de utilidad para mejorar la precisión del test. No obstante, aún resulta necesario solventar los problemas mencionados así como otros de orden práctico que surgen al crear el test definitivo. Por ejemplo, si se requiere la misma di-

ficultad varias veces, el generador de ítems producirá varias veces el mismo ítem salvo que se incluyan controles para limitar la tasa de exposición. Estos controles sobre la tasa de exposición no solo deben afectar a los ítems, sino que es necesario extenderlos a su contenido, logrando la mayor variedad posible en las instrucciones y figuras que se aplican a un mismo sujeto y en el total de los tests. En la elaboración del test definitivo también es necesario tener en cuenta el tiempo que se concede a los sujetos para completar la prueba. Los parámetros han si-

do calibrados concediendo a los sujetos 20 minutos para responder a 50 ítems, por tanto debe imponerse un límite de tiempo similar en el test final para que la rapidez con que deben responderse los ítems no provoque un cambio de sus cualidades psicométricas.

#### Agradecimientos

Esta investigación ha sido financiada, en parte, por los proyectos DGICYT PS94-0040 y PS95-0046.

#### Referencias

- Assessment Systems Corporation (1994). *User's manual for the MicroCAT Testing System, Version 3.5*. St. Paul, MN: Author.
- Bejar, I. 1980. A procedure for investigating the unidimensionality of achievement tests based on item parameters estimates. *Journal of Educational Measurement*. 17, 283-296.
- Bejar, I. 1993. A Generative Approach to Psychological and Educational Measurement. En N. Frederiksen, J. R. Mislevy, R.J. y Bejar I. (eds.). *Test Theory for a New Generation of Tests*. Lawrence Erlbaum Associates. Hillsdale, NJ.
- Brookshear, J. G. 1989. *Theory of Computation. Formal Languages, Automata and Complexity*. The Benjamin/Cummings Publishing Company, Inc. Reedwood City. CAL.
- Collis, J. M.; Tapsfield, P. G. C.; Irvine, S. H.; Dann, P. L. and Wright D. 1995. The British Army Recruit Battery Goes Operational: from Theory to Practice in Computer-Based Testing Using Item-Generation Techniques. *International Journal of Selection and Assessment*. Vol 3. No 2. pp 96-104.
- Embretson, S. E. 1994. Applications of Cognitive Design Systems to Test Development. En C. R. Reynolds, *Cognitive Assessment. A Multidisciplinary Perspective*. Ed Plenum Press. New York.
- Fischer, G. H. y Molenaar, I. W. 1995. *Rasch Models. Foundations, Recent Developments and Applications*. Ed: Springer-Verlag. New York.
- Fischer, G. H. y Pendl, P. 1980. Individualized Testing on the Basis of the Dichotomous Rasch Model. En: L. J. van der Kamp, W. F. Langerak y D. N. de Gruijter (eds.), *Psychometrics for Educational Debates*. Ed: John Wiley & Sons. New York.
- Fischer G.H. y Ponocny-Seliger, E. (1997). LPCM-WIN Program. IEE ProGAMMA. Groningen.
- Hambleton, R. K. y Swaminathan. 1985. *Item Response Theory. Principles and Applications*. Kluwer Nijhoff Publishing. Boston. MA.
- Hornke, L. F. y Habon, M. W. 1986. Rule Based Item Bank Construction and Evaluation with the Linear Logistic Framework. *Applied Psychological Measurement*. Vol 10. No 4. pp 369-380.
- Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates. Hillsdale, NJ.
- SHL. 1996. *DA5: Diagramas Codificados*. SHL, Psicólogos Organizacionales. Madrid.
- Snow R. and Lohman D. 1993. Cognitive Psychology, New Test Design and New Test Theory: An Introduction. In N. Frederiksen, J. R. Mislevy e I. I. Bejar (eds. ), *Test Theory for a New Generation of Tests*. Lawrence Erlbaum Associates. Hillsdale, NJ.
- Wainer, H. 1990. *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates. Hillsdale, NJ.
- Wright, B. D. y Stone, M. H. 1979. *Best test design*. Mesa Press. Chicago.

Aceptado el 6 de abril de 1998