Psicothema
2000. Vol. 12, Supl. nº 2, pp. 198-202

ISSN 0214 - 9915 CODEN PSOTEG

Desarrollos didácticos y funcionamiento diferencial de los items. Problemas inherentes a toda investigación empírica sobre sesgo

Paula Elosua Oliden, Alicia López Jáuregui y Esther Torres Álvarez
Universidad del País Vasco

En este trabajo se exponen los problemas que surgen en toda investigación empírica sobre sesgo. Por un lado la inexistencia de una concordancia absoluta entre los distintos procedimientos de detección del funcionamiento diferencial del ítem (FDI) y por otro la falta de correspondencia entre el concepto de funcionamiento diferencial del ítem y el sesgo. Estos problemas se analizan a través del estudio del FDI en un test de aptitud numérica. Se estudian las fuentes de sesgo producidas por el desarrollo cognitivo y la proximidad temporal entre la instrucción y la administración de la prueba. Se comparan los resultados obtenidos por el estadístico Mantel-Haenszel y el χ^2 de Lord (modelo logístico de 2 parámetros y modelo logístico de 3 parámetros). Los resultados evidencian la disparidad de conclusiones a las que se puede llegar en función del procedimiento de detección del FDI utilizado.

Didactic developments and differential item functioning. This paper sets out the problems that arise in any empirical research on bias: on the one hand, tha fact that the different procedures used to detect the differential item functioning (DIF) do not match exactly, and on the other hand, a lack of correspondence between the concept of DIF and the bias. These problems were analysed by studying the differential item functioning in a numerical ability tests. We examined the source of bias originated by the cognitive development and the interval between instruction and administration of the test. The results obtained by Mantel-Haenszel statistic and the Lord χ^2 procedure (under 2 and 3 parameter logistic models) were compared. Results show tha different conclusions can be drawn according to the procedure used to detect DIF.

El análisis de la validez es una fase imprescindible en el proceso de construcción de instrumentos de medición psicopedagógicos, en la que se recogen evidencias para la confirmación de las hipótesis postuladas respecto a la variable medida, y para la justificación de las inferencias basadas en las puntuaciones obtenidas (Cronbach, 1971). Es un proceso continuo de acumulación de experiencias referidas a situaciones y aplicaciones específicas que exige análisis tanto lógicos y como empíricos.

Si aceptamos la definición de sesgo como error sistemático que distorsiona el significado de las puntuaciones y que está causado por la intervención de habilidades espurias junto a la habilidad principal (Ackerman, 1992; Mellenbergh, 1989; Shealy y Stout, 1993), podemos incluir su evaluación dentro del análisis de la validez. Desde esta visión integradora, el estudio del sesgo al igual que el de la validez se convierte en un proceso continuo de recogida de información en poblaciones concretas y usos determinados.

La posibilidad de que dentro del grupo destinatario de una prueba de medición psicopedagógica sea posible definir subgrupos en función de variables tales como el sexo, la edad, o la experiencia instruccional recibida, que involuntaria e inadvertidamente incorporen un factor contaminante al proceso de medición, obliga a que en los procesos de validación se incluyan evaluaciones del posible sesgo. Estas pueden comenzar con la aplicación de técnicas estadísticas para la detección del funcionamiento diferencial de los ítems. La confirmación de funcionamiento diferencial sin embargo, ha de interpretarse con cautela. El rechazo estadístico de la hipótesis nula implica tan solo la aceptación de un *sesgo hipotético* en el que será menester profundizar con un análisis de contenido y contexto, dirigido por expertos en el área evaluada. Solo así podrá determinarse la existencia o no de interacción entre el contenido del ítem y alguna característica del grupo que contamine el proceso de medida.

En el estudio del sesgo por tanto, al igual que en el análisis de la validez, es necesaria la utilización conjunta de procedimientos estadísticos y lógicos que complementen y refuercen sus resultados (Hambleton, Clauser, Mazor y Jones, 1993; Scheuneman, 1987). Serían dos las fuentes de variación a analizar, el sujeto y el ítem (Mellenbergh y Kok, 1991). En el primer caso el objetivo se centraría en el análisis de los rasgos o variables que operan en el sujeto y condicionan sus respuestas, en el segundo caso se evaluarían las habilidades medidas.

En este contexto general del estudio del sesgo, este trabajo tiene como mayor objetivo exponer o discutir los problemas con que cuenta toda investigación empírica debido sobre todo a la falta de univocidad entre el concepto de funcionamiento diferencial y sesgo. Para ello se analizan las posibles fuentes de sesgo que puede encontrarse en una prueba de aptitud numérica diseñada para cu-

Correspondencia: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20009 San Sebastián (Spain)
E-mail: pspelolp@sc.ehu.es

brir un rango de edad que cubre más de un curso académico. El desarrollo de los componentes cognitivos (Mayer, 1985) implicados en la resolución de problemas de enunciado, así como la proximidad entre la instrucción y la administración de la prueba de evaluación, son factores que pueden contaminar el proceso de medida, distorsionando así los resultados obtenidos.

Método

Sujetos

La muestra está formada por 356 niños con edades comprendidas entre los 9 y los 11 años que estudian en los cursos 4º (N=211) y 6º (N=145) de enseñanza primaria. De ellos 139 pertenecen a un centro de enseñanza público y los 217 restantes a un centro privado concertado de Vitoria-Gasteiz.

Los datos provienen de la administración de una prueba de aptitud numérica aplicada en mayo del curso escolar 1994-95 por una persona especialmente instruida para ello. El test pertenece a la Batería de Aptitudes Diferenciales y Generales en su versión elemental (BADYG-E) (Yuste, 1988). Consta de 25 ítems de elección múltiple con 5 alternativas de respuesta. El coeficiente de fiabilidad aportado por el autor y calculado por el método de dos mitades con la corrección de Spearman-Brown es de 0,86 para 4º curso; el manual no incorpora la información correspondiente a 6º curso, ni los índices de consistencia interna para cada uno de los niveles.

Evaluación de la unidimensionalidad

La unidimensionalidad se evalúa con dos procedimientos. Uno tradicional basado en la varianza explicada por el primer factor tras someter a un análisis de componentes principales la matriz de correlaciones tetracóricas. El otro, el DIMTEST (Stout, 1987; Nandakumar y Stout, 1993) es un procedimiento no paramétrico diseñado para el análisis de la dimensionalidad esencial de datos binarios, que se muestra eficaz en los estudios en los que se ha utilizado (Elosua, López y Egaña, en prensa; Hattie, Krakowski, Rogers y Swaminathan, 1996; Nandakumar, 1994; Nandakumar y Yu, 1996; Padilla, Pérez y González, 1999).

Funcionamiento diferencial de los ítems

La definición más general de funcionamiento diferencial del ítem podría ser la aportada por Mellenbergh (1989), según la cual dada una variable Z, y con respecto a otra variable G, el ítem *i* carece de funcionamiento diferencial, si y sólo si, se satisface la siguiente igualdad para todos los valores *g* y *z* de las variables G y Z.

$$f(X \mid g,z) = f(X \mid z)$$

El carácter de la variable condicionante (observada o latente) permite la clasificación de las técnicas de detección del FDI en dos grupos englobados bajo los epígrafes generales de invarianza condicional observada e invarianza condicional latente (Millsap y Everson, 1993).

Los procedimientos incluidos en el primer grupo definen la variable condicionante Z como la puntuación total observada obtenida por cada sujeto en la prueba. Dentro de este apartado general, se encuadran los procedimientos chi-cuadrado (Scheuneman, 1979), el estadístico Mantel-Haenszel (Holland y Thayer 1988), la estandarización (Dorans y Kulick, 1986), los modelos log-lineales (Mellenbergh, 1982) y los derivados de la regresión logística (Swaminathan y Rogers, 1990).

Dentro del segundo conjunto pueden incluirse todos los procedimientos derivados de la aplicación del modelo de medida propuesto por la teoría de respuesta al ítem (TRI). Dentro de éstos, podemos encontrar procedimientos que comparan los parámetros de las curvas características del ítem (a, b, c) (Lord, 1977, 1980; Mellenbergh, 1982; Wright, Mead y Draba, 1976), y otros que se basan en el cálculo de la superficie que limitan las curvas características producidas por un ítem en dos poblaciones distintas (Linn y Harnisch. 1981; Rudner, 1977; Shepard, Camilli y Williams, 1985; Kim y Cohen, 1991; Raju, 1988, 1990).

En esta investigación se comparan los resultados de la aplicación de dos procedimientos de detección del FDI pertenecientes a cada uno de los grupos, el estadístico Mantel-Haenszel (Holland y Thayer, 1988) perteneciente al grupo de invarianza condicional observada y el chi-cuadrado de Lord (1980) (Invarianza condicional latente).

La aplicación del procedimiento Mantel-Haenszel se lleva a cabo con el programa MHDIF (Fidalgo, 1994), que permite la detección del funcionamiento diferencial del ítem tanto uniforme como no uniforme (Mazor, Clauser y Hambleton, 1994) e incorpora un procedimiento de purificación del criterio en dos etapas.

Para la comparación de los parámetros de los ítems aplicamos el procedimiento ideado por Lord (1980). Para mejorar su efectividad seguimos las pautas aconsejadas por Candell y Drasgow (1988). Una vez estimados los parámetros en cada grupo, y dada la arbitrariedad de la escala de θ, se equiparan las escalas y se estima el FDI. En una segunda fase se eliminan los ítems con FDI y se reequiparan las escalas, volviendo a detectar el FDI sobre todos los ítems. Este procedimiento se ejecuta una y otra vez hasta que en dos iteraciones consecutivas los resultados sean coincidentes. La equiparación de las escalas se lleva a cabo por el método de la *curva característica* (Stocking y Lord, 1983) implementado en el programa EQUATE (Baker, 1994) y el análisis del funcionamiento diferencial con IRTDIF (Kim y Cohen, 1992).

Resultados

Los primeros estadísticos descriptivos muestran que la media aritmética del grupo 6º ($\bar{X}$=17,94; S_X=4,00) es mayor que la obtenido por el grupo 4º ($\bar{X}$=16,65 ;S_X=4,19), siendo la diferencia entre ellos significativa (t=-2,902; p=0,004). La consistencia interna se evalúa con el alpha de Cronbach (1951), que arroja los valores de 0,806 para 4º y 0,788 para 6º.

Tabla 1
Porcentaje de varianza explicada por los factores

	4º		6º	
factores	**Valor propio**	**%Varianza explicada**	**Valor propio**	**%Varianza explicada**
1	6.588	26.34%	9.507	38.02%
2	2.641	10.56%	3.580	14.32%
3	1.759	7.03%	2.899	11.59%

Evaluación de la unidimensionalidad

Los porcentajes de varianza explicada por los 3 primeros factores en cada una de las muestras puede observarse en la tabla 1. Para las dos muestras se supera el tan utilizado criterio de unidimensionalidad de Reckase (1979).

En la aplicación del DIMTEST el subtest AT1 se forma con 5 ítems que selecciona automáticamente el programa de los resultados del análisis de componentes principales ejecutado a partir de la matriz de correlaciones tetracóricas. Los dos conjuntos de datos superan el test de Wilcoxon que contrasta la hipótesis de que los ítems seleccionados no sean excesivamente fáciles. Para 4º y 6º los valores de *p* son correlativamente, p=0,06 y p=0,227.

La tabla 2 recoge los resultados de la aplicación de este procedimiento sobre cada uno de los grupos de datos. Puede verse que en los dos casos se acepta la hipótesis contrastada de unidimensionalidad esencial.

Tabla 2
Evaluación de la unidimensionalidad esencial

	T conservador				T más potente			
	T_1	**T_2**	**T**	**p**	**T_1**	**T_2**	**T**	**p**
4	-0,1290	-0,9403	0,5736	0,283	-0,2313	-1,2676	0,7328	0,2318
6	-0,9605	-0,3373	-0,4406	0,6702	-1,3500	-0,4898	-0,6082	0,7284

Tabla 3
Funcionamiento diferencial de los ítems (**p<0,01)

Item	χ^2 Lord tres parámetros	Δ_{MH}	Item	χ^2 Lord tres parámetros	Δ_{MH}
1	3.7348	-0.68	1	0.3900	-0.68
2	0.1511	0.06	2	0.1100	0.06
3	9.5818	1.18	3	1.7640	1.18
4	2.5048	-1.39	4	2.8816	-1.39
5	1.6310	1.60	5	2.9840	1.60
6	1.8464	0.09	6	1.1343	0.09
7	2.9476	-2.55	7	0.9692	-2.55
8	0.7009	0.15	8	0.2208	0.15
9	4.9089	0.58	9	1.9215	0.58
10	0.4117	-2.07	10	1.5295	-2.07
11	4.4776	-0.29	11	0.0384	-0.29
12	7.3020	0.60	12	0.8928	0.60
13	1.6757	-0.10	13	0.0944	-0.10
14	**12.2410****	-0.09	14	1.9887	-0.09
15	**11.9587****	-1.15	15	1.5913	-1.15
16	**14.3963****	-0.74	16	1.6061	-0.74
17	3.0371	-0.34	17	0.8923	-0.34
18	1.2586	0.57	18	1.2404	0.57
19	2.0761	-0.76	19	0.4901	-0.76
20	6.5721	**2.16****	20	**16,3285****	**2.16****
21	10.3811	0.58	21	8.9446	0.58
22	1.8948	0.84	22	8.1643	0.84
23	**29.5548****	**2.84****	23	**14.5029****	**2.84****
24	**17.6515****	**4.04****	24	**49.1746****	**4.04****
25	0.8299	0.17	25	1.0215	0.17
	5	3		3	3
CONCORDANCIAS 2			**CONCORDANCIAS 3**		

Estimación de los parámetros

Dado el objetivo primero de este trabajo, mostrar los problemas con los que cuenta la investigación empírica sobre FDI, se estiman los parámetros de los ítems con dos modelos; el logístico de dos parámetros y el logístico de tres parámetros. El procedimiento de estimación utilizado en ambos casos, es el implementado en BILOG (Mislevy y Bock, 1990), la estimación marginal por máxima verosimilitud.

En ninguno de los cursos analizados aparecen ítems con valores chi-cuadrado de ajuste significativos (p<0,01).

Funcionamiento diferencial de los ítems

Aplicadas las técnicas de detección del funcionamiento diferencial de los ítems los resultados pueden verse en la tabla 3, dónde los asteriscos corresponden a índices de FDI significativos (p<0,01).

Bajo el epígrafe χ^2 aparecen los valores de este estadístico, tras un proceso iterativo de purificación del criterio que converge en dos etapas para cada uno de los modelos utilizados, y cuyas constantes de equiparación son en el modelo logístico de dos parámetros , A_1=0,8927 - K_1=—0,3548 y A_2=0,8845 - K_2=-0,5373, y en el modelo logístico de tres parámetros A_1=0,9010 - K_1=-0,2889 y A_2=0,8058 - K_2=-0,3851. La columna Δ_{MH} muestra los índices delta tras un proceso bietápico de purificación de la puntuación observada. Los valores positivos de Δ_{MH} indican funcionamiento diferencial a favor del grupo focal, que en nuestro caso es el formado por los sujetos de 4º.

Una vez aplicada la corrección propuesta por Mazor, Clauser y Hambleton (1994), para la detección del FDI no uniforme, los valores obtenidos son significativos para los ítems 20 y 24 en el grupo de puntuación inferior (Δ_{MH20}=2,81 y Δ_{MH24} =7,55). Entre los sujetos que obtienen puntuaciones más elevadas presentan funcionamiento diferencial los ítems 23 y 24 para los que el valor Δ_{MH} es de 2,84.

Es de reseñar que estos 3 ítems 20, 23 y 24 han sido también detectados en el análisis del funcionamiento diferencial uniforme, y es de destacar el hecho de que los tres favorecen al grupo de 4º

Como resultado de la aplicación de estas técnicas de detección puede verse que el estadístico Mantel-Haenszel cataloga 3 ítems con funcionamiento diferencial (20, 23 y 24) mientras que el chi-

Tabla 4
Análisis de contenido y funcionamiento diferencial de los ítems en función de los procedimientos de detección utilizados

Contenido		**Nº ítems**	**Items con FDI 2 parámetros -M.H**	**Items con FDI 3 parámetros -M.H.**
Items aritméticos		11	Ninguno	Ninguno
Items de enunciado	Sin conocimientos específicos	8	Ninguno	14-15-16 (6º)
	Con conocimientos específicos	4	23(4º)	23(4º)
Items declarativos		2	20-24 (4)	20-24(4º)
Total		25	3	6

cuadrado de Lord detecta 5 ítems bajo el modelo logístico de 3 parámetros (14,15,16,23 y 24) y tres ítems (20, 23 y 24) en el modelo logístico de dos parámetros.

El número de coincidencias en la detección de ítems con funcionamiento diferencial depende por tanto del modelo de respuesta al ítem utilizado. En el caso de que el modelo no incorpore el parámetro de pseudo-azar la correspondencia entre MH y Lord es del 100%, mientras que la incorporación de este parámetro reduce el nivel de acuerdo entre ambas técnicas. La consideración del estadístico Mantel Haenszel como una extensión del modelo logístico de un parámetro (Hambleton y Rogers 1989; Holland y Thayer, 1988; Thissen y Steinberg, 1988), explicaría el hecho de que la incorporación del parámetro de pseudo-azar al modelo distanciara los resultados obtenidos por ambos métodos de detección. Estos resultados evidencian un alto nivel de concordancia entre ambos procedimientos atestiguada en otras investigaciones (Hambleton y Rogers, 1989; Raju, Drasgow y Slinde, 1993; Elosua, López y Egaña, en prensa) que se ve aminorada con la incorporación del parámetro de pseudo-azar al modelo, circunstancia ésta encontrada en trabajos anteriores (Elosua, López , Egaña, Artamendi y Yenes, en prensa).

Conclusiones

El hecho de que los índices de funcionamiento diferencial obtenidos tras la aplicación de técnicas más o menos complejas no se puedan considerar necesariamente pruebas de sesgo contra uno de los grupos obliga a que a todo análisis cuantitativo siga uno cualitativo que determine tras una revisión de contenido la presencia o no de errores sistemáticos de medida que amenacen la validez del instrumento en cuestión.

Este tipo de análisis se lleva a cabo con los ítems que muestran valores positivos en el estudio exploratorio anterior. El problema en este punto viene dado por la falta de concordancia absoluta entre los distintos procedimientos de detección de FDI que podemos encontrar en la literatura especializada. Es más, como mostramos en este trabajo, un mismo procedimiento puede arrojar resultados diferentes en función del modelo de medida en que se aplique. Esta indeterminación deja en manos del investigador el posterior análisis de los ítems.

En nuestro caso, un análisis de contenido de los ítems que componen la prueba nos lleva a categorizarlos en tres grupos generales:

– Items aritméticos: 11 ítems sin contenido verbal, en los que se exige la ejecución de una de las cuatro operaciones aritméticas básicas. Según el modelo propuesto por Mayer exigirían únicamente el proceso de ejecución y los correspondientes conocimientos algorítmicos (ítems 1-2-3-4-5-6-8-10-11-17-19).

– Doce ítems que implican la resolución de problemas de enunciado. Ocho de los cuales requieren la realización de alguna o algunas de las operaciones aritméticas básicas (ítems 7-9-13-14-15-16-18), mientras que para la resolución de los restantes ítems son precisos de modo adicional conocimientos factuales de equivalencias de fracciones (ítem 21), magnitudes (ítem 23) y geometría (ítems 22-25).

– Dos ítems que requieren exclusivamente de conocimientos factuales de tipo *declarativo* (equivalencia de magnitudes y escritura de números romanos) en los que no es preciso ejecutar operación aritmética alguna (ítems 20-24).

Aunando los resultados obtenidos tras la aplicación de los procedimientos de detección de FDI, y el análisis de contenido, llegamos a los resultados mostrados en la tabla 4.

1. Ausencia de FDI en todo el bloque de ítems correspondientes a operaciones aritméticas

2. Presencia de funcionamiento diferencial a favor del curso superior en tres de los problemas de enunciado (ítems 14-15-16) cuando el modelo de TRI utilizado es el logístico de tres parámetros. FDI encontrado por todos los procedimientos de detección a favor del grupo inferior en un problema de enunciado con conocimiento factual (ítem 23).

3. Funcionamiento diferencial que favorece al grupo inferior en los dos ítems declarativos (ítems 20 y 24). En este caso los resultados obtenidos con el chi-cuadrado de Lord son divergentes en función del modelo logístico en el ítem 20, si bien en referencia al 24 existe unanimidad entre las técnicas utilizadas.

La búsqueda de conclusiones por tanto, esta limitada por el modelo utilizado. Bajo el modelo logístico de tres parámetros se puede mantener la hipótesis de la influencia del desarrollo cognitivo en la resolución de problemas de enunciado. Hipótesis por otro lado, que no se sostiene si sólo se contrastaran los parámetros de dificultad y discriminación.

La ruptura o falta de univocidad entre el concepto estadístico de funcionamiento diferencial del ítem y el sesgo en la medida, puede verse como un problema recurrente dentro de la investigación psicológica. La falta de correspondencia entre las definiciones operacionales y teóricas de los constructos expuesta de modo ejemplar por Torgerson (1958) vuelve a repetirse en el estudio del sesgo. Del mismo modo, las salidas posibles por parte del investigador a este problema, rechazo de las hipótesis de partida o rechazo de los indicadores del constructo utilizados, vuelven a ser plausibles en esta área de investigación. Ante circunstancias como las expuestas y siempre en función de los intereses tanto del investigador como de la investigación se presentan dos alternativas, o bien rechazar las hipótesis sobre sesgo o bien aceptarlas en función de los intereses últimos de la investigación.

Referencias

Ackerman, T.A.(1992). Didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Baker, F.B. (1994) EQUATE2: Computer program for equating two metrics in item response theory [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design

Candell, G.L. y Drasgow, F.(1988): An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement,* 12(3), 253-260.

Cronbach, L.J.(1951). Coefficient Alpha and the Interntal Structure of Tests, *Psychometrika, 16*, 297-334.

Cronbach, L.J.(1971). Test validation. In R.L. Thorndike(Ed.), *Educational Measurement*(pp.443-507). Washington, DC: American Council on Education.

Dorans, N.J. y Kulick, E.(1986). Demonstrating the utility of the Standarization Approach to assessing unexpected Differential Item Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368.

Elosua, P. López, A y Egaña, J. (en prensa) Idioma de aplicación y rendimiento en una prueba de comprensión verbal.

Elosua, P. López, A y Egaña, J. (en prensa) Fuentes potenciales de sesgo en una prueba de aptitud numérica.

Elosua, P., López, A., Egaña, J., Artamendi, J. y Yenes, F. (en prensa) Funcionamiento diferencial de los ítems en la aplicación de pruebas psicológicas en entornos bilingües.

Fidalgo, A.M.(1994). MHDIF: A computer program for detecting uniform and nouniform differentzial item functioning with the Mantel-Haenszel procedure.[Computer program] Dpto. Psicología, Universidad de Oviedo.

Hambleton, R.K., Clauser, B.E., Mazor, K.M. y Jones, R.W. (1993) Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Hambleton, R.K. y Rogers, H.J.(1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel Methods. *Applied Measurement in Education, 2*(4), 313-334.

Hattie, J., Krakowski, K., Rogers, H.J. y Swaminathan, H.(1996) An assessment of Stout's index of essential unidimensionality. *Applied psychological measurement, 20*(1), 1-14.

Holland, P.W. y Thayer, D.T.(1988). Differential Item Performance and the Mantel-Haenszel procedure. En H. Wainer y H.J. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Kim, S.H. y Cohen, A.S.(1991). A comparison of two area measures for detecting Differential Item Functioning. *Applied Psychological Measurement, 15*(3), 269-278.

Kim S.H. y Cohen, A.S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis [Computer Program] University of Wisconsin-Madison.

Linn, R.L. y Harnisch, D.L.(1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109-118.

Lord, F.M.(1977). A study of item bias, using item characteristic curve theory. En Y.H. Poortinga(Ed.), *Basic problems Cross-Cultural Psychology* (pp.19-29).Amsterdam: Swets y Zeitlinger.

Lord, F.M.(1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Mayer, R.E. (1985). Capacidad matemática. En R.J. Sternberg, (de.) *Human abilities. An information processing approach.* New York: Freeman and company. (Trad. Cast. Las capacidades humanas. Un enfoque desde el procesamiento de la información. Barcelona. Labor, 1986)

Mazor, K.M., Clauser, P.E. y Hambleton, R.K. (1994). Identification of nonuniform Differential Item Functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*, 284-291.

Mellenbergh, G.J.(1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118.

Mellenbergh, G.J.(1989). Item bias and Item Response Theory. *International Journal of Educational Research, 13*, 127-143.

Mellenbergh, G.J. y Kok, F.G.(1991). Finding the biasing trait(s). In P.L. Dann, S.H. Irvine y J.M. Collins(Eds.), *Advances in computer-based human assessment*(pp.291-306). Dordrecht: Kluver Academic Publishers.

Millsap, R.E. y Everson, H.T.(1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Mislevy, R.J. y Bock, R.D.(1990). BILOG·3: Item analysis and test scoring with binary logistic models.[Computer program]. Mooresville, IN: Scientific software.

Nandakumar, R. (1994) Assessing dimensionality pf a set of item responses. Comparison of different approaches. *Journal of educational measurement, 31*, 17-35.

Nandakumar, R. y Stout, W. (1993) Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of educational statistics, 18*,41-68

Nandakumar, R. y Yu, F. (1996) Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of educational measurement, 33*, 355-368.

Padilla, J.L., Pérez, C. y González, A. (1999) Efecto de la instrucción sobre la dimensionalidad del test. *Psicothema, 11*(1), 183-193.

Raju, N.S.(1988). The area between two item characteristic curves. *Psychometrika,53*(4), 495-502.

Raju, N.S.(1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Raju, N.S., Drasgow, F. y Slinde, J.A.(1993). An empirical comparison of the area methods, Lord's Chi-square test and Mantel-Haenszel technique for assessing Differential Item Functioning. *Educational and Psychological Measurement, 53*(2), 301-314.

Reckase, M.D.(1979). Unifactor latent trait models applied to multifactor test: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Rudner, L.M.(1977, April). *An approach to biased item identification using latent trait measurement theory.* Paper presented at the Annual Meeting of The American Educational Research Association, New York.

Shealy, R. y Stout, W.(1993). An Item Response Theory model of test bias and Differential Test Functioning. In W.P. Holland y H. Wainer(Eds.), *Differential Item Functioning*(pp.197-240). Hillsadale, NJ: Lawrence Erlbaum.

Shepard, L.A., Camilli, G. y Williams, D.M.(1985). Validity of aproximation techniques for detecting item bias. *Journal of Educational Measurement, 22*(2), 77-105.

Scheuneman, J.D.(1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*(3), 143-152.

Scheuneman, J.D.(1987). An experimental exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24*(1), 97-118.

Stocking, M.L. y Lord, F.M. (1983) Developing a common metric in item response theory. *Applied psychological measurement, 7*(2), 201.210.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Swaminathan, H. y Rogers, H.J.(1990). Detecting Differential Item Functioning using logistic regresion procedures. *Journal of Educational measurement, 27*(4), 361-370.

Thissen, D. y Steinberg, L.(1988). Data analysis using Item Response Theory. *Psychological Bulletin, 104*(3), 385-395.

Torgerson, W.S.(1958). *Theory and methods of scaling.* New York: John Wiley.

Wright, B.D., Mead, R. y Draba, R.(1976). *Detecting and correcting item bias with a logistic response model.*(Research memorandum, Nº22). Chicago, IL: University of Chicago, Statistical lab., Departament of Education.

Yuste, C. (1988). BADYG-E. Madrid. Ciencias de la educación preescolar y especial.