Psicothema, 1994. Vol 6, n.º 3, pp. 493-501
ISSN 0214 - 9915 CODEN PSOTEG

UTILIZACION DEL ANALISIS FACTORIAL Y MEDIDAS DEL AREA COMO METODOS EN LA DETECCION DE SESGO[1]

María José Navas Ara
Universidad Nacional de Educación a Distancia

El presente trabajo tiene por objeto examinar la posible existencia de sesgo respecto al sexo en dos pruebas muy utilizadas en el ámbito industrial y educativo: el factor numérico de la batería de Tests de Aptitudes Escolares (nivel 3) y el de la batería de Tests de Aptitudes Diferenciales. Para estudiar el sesgo se han seguido dos aproximaciones distintas: (1) una aproximación factorial, que supone la comparación de las soluciones factoriales obtenidas en varones y mujeres y (2) una aproximación basada en la teoría de respuesta al item, que supone el cálculo de los estadísticos del área –con signo y sin signo– definidos por Raju (1988, 1990). Los resultados obtenidos desde ambas aproximaciones son discrepantes.

Bias Detection Using Factor Analysis and Area Measures. This work aims to examine whether or not there is bias against males and females in two well-known and very widely used tests of numerical ability: those of Tests of Educational Abilities and Differential Aptitude Tests. Two different methods have been used to study bias: (1) a factor analytic method whose purpose was to compare the factorial structures obtained in the male and female samples and (2) an item response theory-based method consisting of calculating the signed and unsigned area statistics (Raju, 1988, 1990). The pattern of results obtained is different depending on the method used.

En la actualidad, el estudio del sesgo en los items y en los tests es un área muy activa de investigación, a pesar de ser un tema tratado tardíamente en la literatura (Muñiz, 1990, 1992) y cuyo estudio ha venido impulsado en gran parte desde fuera del ámbito psicométrico. En este sentido, el gran interés social en el tratamiento igualitario de grupos o minorías étnicas y/o socio-políticas ha sido un factor determinante a la hora de estimular el estudio sobre el sesgo de los items y tests utilizados para la selección de personal o en la promoción y orientación educativas. El sesgo de los instrumentos de medida se ha revelado como algo más que una cuestión puramente técnica, convirtiéndose en una cuestión de debate público e incluso legal. Title (1988) afirma precisamente que la tensión entre la práctica profesional de la evaluación y la presión de la opinión pública ha dado como resultado una serie de estudios que requieren prestar una atención renovada a la teoría que subyace a los tests y a su proceso de desarrollo, así como a una concepción más amplia de la evidencia de la validez necesaria para justificar su uso.

La investigación sobre el sesgo de los items y los tests se ha centrado mayoritariamente en grupos étnicos y, en menor medida, en grupos definidos por el sexo. Desde la

Correspondencia: María José Navas Ara
Facultad de Psicología. Universidad Nacional de Educación a Distancia
Ciudad Universitaria, s/n. 28040. Spain

[1] Una primera versión de este trabajo fue presentada como una comunicación en el III Simposium de Metodología de las Ciencias Sociales y del Comportamiento celebrado en Santiago de Compostela en julio de 1993.

publicación en 1873 de la obra de Herbert Spencer *Psychology of the Sexes* hasta la actualidad, la investigación sobre las diferencias según sexo en habilidad y rendimiento ha sido continua y ha puesto de manifiesto que, generalmente, los varones tienen una mayor habilidad numérica y espacial, mientras que las mujeres tienden a mostrar niveles superiores de rendimiento en estudios lingüísticos y verbales (Anastasi, 1958; Keeves, 1988; Maccoby, 1966; Moss, 1982; Tyler, 1956; Walker, 1976).

El presente trabajo tiene por objeto estudiar la plausibilidad de la hipótesis de diferencias según sexo en la aptitud numérica, examinando para ello la posible existencia de sesgo en dos pruebas comercializadas y muy utilizadas tanto en el ámbito industrial –en la selección de personal– como en el educativo –en la selección, promoción y orientación académicas: el factor numérico de la batería de Tests de Aptitudes Escolares nivel 3 (FNTEA) y el de la batería de Tests de Aptitudes Diferenciales (FNDAT). En definitiva, se trataría de ver si existe impacto –diferencias en la actuación en la prueba debidas a diferencias reales en la habilidad numérica– o, por el contrario, sesgo –funcionamiento diferencial en los items de estas dos pruebas en varones y mujeres.

Método

Las dos pruebas consideradas –FNTEA y FNDAT– miden la aptitud numérica a través de una serie de ejercicios de cálculo numérico que implican operaciones algebraicas básicas con enteros, decimales, fracciones, cálculo de porcentajes, etc.. Constan respectivamente de 30 y 40 items de elección múltiple, con cinco alternativas de respuesta cada uno de ellos.

La prueba FNTEA fue aplicada a una muestra de 9.058 alumnos –4.913 varones y 3.825 mujeres– de 2.° curso de Enseñanzas Medias (EE.MM.) y la prueba FNDAT a una muestra de 7.650 alumnos –3.590 varones y 3.780 mujeres– de 4.° curso de EE.MM. (Curso de Orientación Universitaria ó 2.° curso del 2.° grado de Formación Profesional).

Para abordar el estudio del sesgo de estas dos pruebas se han seguido dos aproximaciones diferentes: una aproximación factorial y una aproximación basada en la Teoría de Respuesta al Item (TRI).

La aproximación factorial supone realizar sendos análisis factoriales en las muestras de varones y mujeres y comparar las soluciones factoriales obtenidas en ambos casos. El método de extracción de factores utilizado fue el de factores principales y no se llevó a cabo ningún tipo de rotación.

Para comparar las soluciones factoriales obtenidas en la muestra de varones y en la de mujeres, se llevó a cabo el test de la congruencia (Pine, 1977; Rummel, 1970):

$$C_{vm} = \sqrt{\frac{\sum_{i=1}^{n}(r_{iv} - r_{im})^2}{n}} \qquad (1)$$

donde:

C_{vm}: coeficiente de congruencia entre varones y mujeres

r_{iv}: peso factorial en el primer factor del item i obtenido en la muestra de varones

r_{im}: peso factorial en el primer factor del item i obtenido en la muestra de mujeres

n: número de items del test

y se calculó también el índice de congruencia de Burt y Tucker:

La aproximación basada en la TRI exige como paso previo al estudio del sesgo la

$$IC_{vm} = \frac{\sum_{i=1}^{n} r_{iv} r_{im}}{\sqrt{\left(\sum_{i=1}^{n} r_{iv}^2\right)\left(\sum_{i=1}^{n} r_{im}^2\right)}} \qquad (2)$$

comprobación del ajuste de los datos a alguno de los modelos de la TRI. Para ello, se examinaron tres supuestos distintos:

i) el supuesto de unidimensionalidad, mediante el test del autovalor y el test de la línea base aleatoria.

Se llevaron a cabo análisis factoriales sucesivos en la matriz global de datos hasta obtener una estructura factorial en la que todos los items saturaban en el primer factor con saturaciones factoriales superiores a 0.25. El método de extracción de factores fue el de factores principales, no se llevó a cabo ningún tipo de rotación y se forzó el análisis a nueve factores.

ii) el supuesto de ausencia de aciertos al azar, mediante el estudio de la actuación de los sujetos con puntuaciones más bajas en el test en los items más difíciles.

iii) el supuesto de discriminación constante, mediante el estudio de la distribución de frecuencias de la correlación biserial puntual entre las puntuaciones de cada item y la puntuación total del test.

Para examinar la existencia de sesgo se han calculado los estadísticos del área con signo y sin signo (Raju, 1988, 1990) entre las curvas características de los items estimadas en la muestra de varones y en la de mujeres, una vez que éstas han sido convenientemente equiparadas:
para el modelo logístico de tres parámetros:

$$EAcs_i = (1 - c_i)\,(b_{iv} - b_{im}) \qquad (3)$$

$$EAss_i = (1 - c_i) \left| \frac{2(a_{iv} - a_{im})}{Da_{iv}a_{im}} \ln\left(1 + \exp\frac{Da_{iv}a_{im}(b_{iv} - b_{im})}{a_{iv} - a_{im}}\right) - (b_{iv} - b_{im}) \right| \qquad (4)$$

donde

$EAcs_i$: estadístico del área con signo del item i

$EAss_i$: estadístico del área sin signo del item i

c_i : valor estimado para el parámetro c del item i

b_{iv}: valor estimado para el parámetro b del item i en el grupo de varones

b_{im}: valor estimado para el parámetro b del item i en el grupo de mujeres

a_{iv}: valor estimado para el parámetro a del item i en el grupo de varones

a_{im}: valor estimado para el parámetro a del item i en el grupo de mujeres

para el modelo logístico de dos parámetros:

$$EAcs_i = (b_{iv} - b_{im}) \qquad (5)$$

$$EAss_i = \left| \frac{2(a_{iv} - a_{im})}{Da_{iv}a_{im}} \ln\left(1 + \exp\frac{Da_{iv}a_{im}(b_{iv} - b_{im})}{a_{iv} - a_{im}}\right) - (b_{iv} - b_{im}) \right| \qquad (6)$$

para el modelo logístico de un parámetro:

$$EAcs_i = (b_{iv} - b_{im}) \qquad (7)$$

$$EAss_i = | b_{iv} - b_{im} | \qquad (8)$$

En el proceso de equiparación realizado el método elegido para poner en la misma escala las estimaciones obtenidas en la muestra de varones y en la de mujeres ha sido el método estándar de la media y la desviación típica (Warm, 1978).

Con el fin de interpretar los resultados obtenidos al calcular dichos estadísticos se procedió a simular la distribución muestral de los mismos. Para ello, se simuló con el DATAGEN (Hambleton y Rovinelli, 1973) una matriz de datos de tamaño igual al número de varones al que se había aplicado cada prueba. Esta matriz se ajustaba perfectamente al mismo modelo al que se ajustaba la matriz original de datos y se generaba dando como parámetros de los items los valores estimados en la matriz global de datos –varones y mujeres– y como parámetros de habilidad de los sujetos los valores estimados en dicha matriz para los varones. Se simuló del mismo modo una matriz de datos de tamaño igual al número de mujeres. En las matrices simuladas se estimaron los parámetros de los items, se equipararon las estimaciones y se calcularon los correspondientes estadísticos del área.

Resultados

Aproximación Factorial

La tabla 1 y los cuadros 1 y 2 recogen los resultados obtenidos al estudiar el sesgo de estas dos pruebas desde la aproximación factorial.

Tabla 1
Resultados obtenidos al estudiar el sesgo desde la aproximación factorial

Prueba	Test de congruencia	Indice de congruencia
FNTEA	0.055	0.995
FNDAT	0.033	0.997

Los valores obtenidos al realizar el test de la congruencia –muy próximos al cero– y al calcular el índice de congruencia –muy próximos a la unidad– apuntan hacia una probable inexistencia de sesgo en ambas pruebas. Este resultado se confirma igualmente al examinar las estructuras factoriales de FNTEA y FNDAT obtenidas en la muestra de varones y en la de mujeres (Véase cuadros 1 y 2).

En efecto, al estudiar ambas pruebas se observa que existe un factor claramente dominante en la muestra de varones y en la de mujeres que explica, en la prueba FNTEA, entre el 78 y el 80% de la variabilidad observada en el espacio factorial y, en la prueba FNDAT, entre el 68 y el 69% de la misma. Además, al examinar los items concretos que saturan en los factores obtenidos se observa que son los mismos en la estructura obtenida en la muestra de varones y de mujeres, para las dos pruebas consideradas y en todos los factores aislados por el análisis.

Aproximación de la TRI

Al examinar el supuesto de unidimensionalidad, los resultados obtenidos ponen de manifiesto que:

i) de los 30 items de la prueba FNTEA sólo los 23 primeros miden una misma dimensión (los 7 restantes definían claramente un factor de velocidad de realización de la prueba)

ii) de los 40 items de la prueba FNDAT sólo 28 miden una misma dimensión (FNDAT9–12, 15, 17, 19–40).

Los gráficos 1 y 2 ilustran los resultados obtenidos con el test del autovalor y con el de la línea base aleatoria en estos items de

Cuadro 1
Estructura Factorial de FNTEA

FACTORES	VARONES			MUJERES		
	AUTOVALORES	% AC. VAR. ESPACIO		AUTOVALORES	% AC. VAR. ESPACIO	
		DATOS	FACTORES		DATOS	FACTORES
1	6.656	.646	.804	5.684	.600	.783
2	1.620	.804	1.000	1.573	.766	1.000

Cuadro 2
Estructura Factorial de FNDAT

FACTORES	VARONES			MUJERES		
	AUTOVALORES	% AC. VAR. ESPACIO		AUTOVALORES	% AC. VAR. ESPACIO	
		DATOS	FACTORES		DATOS	FACTORES
1	7.344	.542	.693	6.900	.530	.684
2	2.135	.700	.895	1.985	.682	.881
3	1.113	.782	1.000	1.206	.775	1.000

las pruebas FNTEA y FNDAT, respectivamente. Estos gráficos presentan los autovalores obtenidos para los nueve factores con la matriz de datos reales y simulados, siendo idéntica la dimensión de ambas (9.058×23 para la prueba FNTEA y 7.650×28 para FNDAT). Como se puede observar en ambos gráficos, el codo de la línea correspondiente al conjunto de datos reales se encuentra en el primer factor y los autovalores de

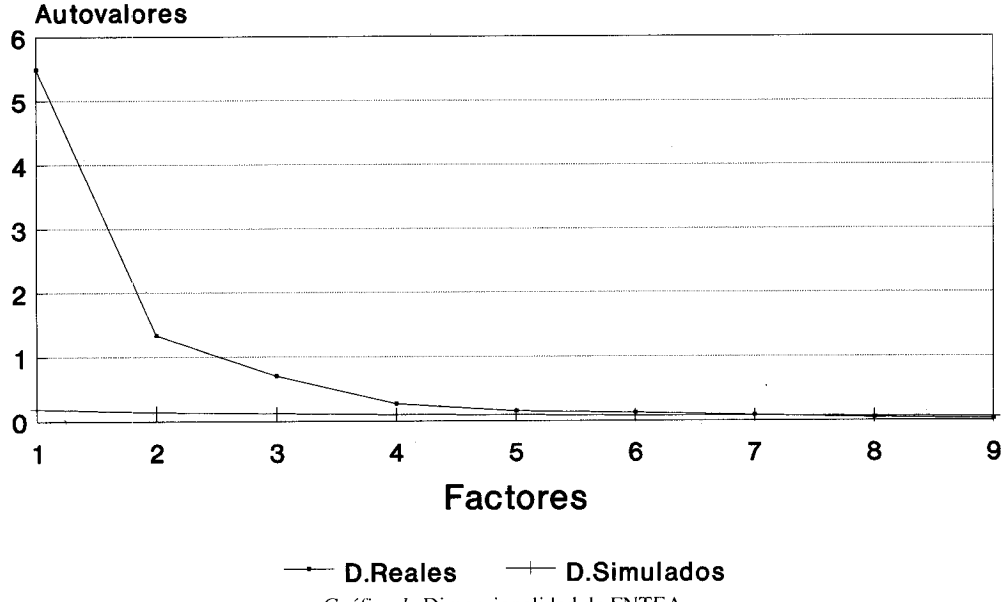

Gráfico 1: Dimensionalidad de FNTEA.

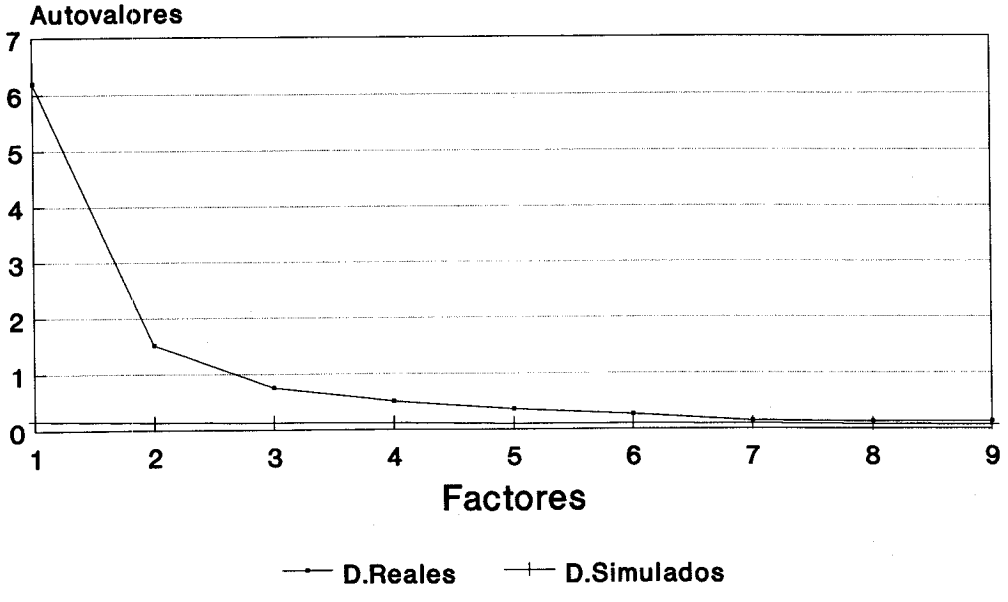

Gráfico 2: Dimensionalidad de FNDAT.

este factor en ambos conjuntos de datos son claramente distintos.

Al examinar los supuestos de ausencia de aciertos al azar y de discriminación constante los resultados obtenidos revelan que en ninguna de las dos pruebas son violados estos dos supuestos, razón por la cual el modelo logístico con el que se va a trabajar es el modelo de un parámetro, el modelo de Rasch.

Las tablas 2 y 3 presentan los resultados obtenidos al estudiar el sesgo desde la aproximación de la TRI para las pruebas FNTEA y FNDAT, respectivamente.

En estas tablas se muestra para cada uno de los items unidimensionales de FNTEA y FNDAT el valor del estadístico del área con signo y sin signo –dos primeras columnas– y los obtenidos al simular la distribución muestral del estadístico del área con y sin signo –dos últimas columnas.

Un criterio plausible para interpretar el estadístico del área puede ser determinar un punto de corte en la distribución muestral obtenida para cada prueba, a partir del cual se consideraría un item como sesgado –si es superior a dicho valor– o insesgado –si es igual o inferior. Ese punto de corte podría ser el valor mayor obtenido en la distribución

Tabla 2
Resultados obtenidos al estudiar el sesgo de FNTEA desde la aproximación de la TRI

ITEM N.º	Estadístico área		Estadístico línea base	
	EAcs	EAss	EAcs	EAss
FNTEA1	.659	.659	–.016	.016
FNTEA2	.712	.712	.046	.046
FNTEA3	.536	.536	.045	.045
FNTEA4	.121	.121	.061	.061
FNTEA5	.397	.397	–.015	.015
FNTEA6	–.476	.476	.007	.007
FNTEA7	.080	.080	–.003	.003
FNTEA8	–.064	.064	–.074	.074
FNTEA9	–.084	.084	.056	.056
FNTEA10	.032	.032	.047	.047
FNTEA11	.146	.146	.015	.015
FNTEA12	.115	.115	–.012	.012
FNTEA13	–.297	.297	–.049	.049
FNTEA14	–.115	.115	.019	.019
FNTEA15	–.075	.075	.010	.010
FNTEA16	–.391	.391	.010	.010
FNTEA17	–.259	.259	–.022	.022
FNTEA18	–.276	.276	–.027	.027
FNTEA19	.140	.140	–.021	.021
FNTEA20	–.069	.069	–.025	.025
FNTEA21	–.235	.235	.010	.010
FNTEA22	–.320	.320	–.012	.012
FNTEA23	–.271	.271	–.056	.056

Tabla 3
Resultados obtenidos al estudiar el sesgo de FNDAT desde la aproximación de la TRI

ITEM N.º	Estadístico área		Estadístico línea base	
	EAcs	EAss	EAcs	EAss
FNDAT9	.123	.123	–.008	.008
FNDAT10	.206	.206	–.017	.017
FNDAT11	.233	.233	–.057	.057
FNDAT12	.220	.220	.021	.021
FNDAT15	.043	.043	.015	.015
FNDAT17	–.351	.351	.005	.005
FNDAT19	–.110	.110	–.090	.090
FNDAT20	.024	.024	–.028	.028
FNDAT21	–.047	.047	–.005	.005
FNDAT22	–.554	.554	.055	.055
FNDAT23	–.086	.086	.037	.037
FNDAT24	.196	.196	.010	.010
FNDAT25	–.223	.223	.026	.026
FNDAT26	–.029	.029	.059	.059
FNDAT27	–.320	.320	.017	.017
FNDAT28	.166	.166	–.007	.007
FNDAT29	.189	.189	–.034	.034
FNDAT30	.183	.183	–.001	.001
FNDAT31	–.109	.109	.012	.012
FNDAT32	–.025	.025	.066	.066
FNDAT33	.042	.042	–.031	.031
FNDAT34	.024	.024	.064	.064
FNDAT35	.131	.131	–.017	.017
FNDAT36	.142	.142	–.058	.058
FNDAT37	–.006	.006	–.005	.005
FNDAT38	.089	.089	–.025	.025
FNDAT39	.177	.177	.004	.004
FNDAT40	–.327	.327	–.007	.007

muestral. En la prueba FNTEA dicho valor sería 0.061 para EAcs y 0.074 para el EAss y en la prueba FNDAT 0.066 para EAcs y 0.090 para el EAss.

Si se examina la tabla 2 se observa que sólo un item –FNTEA10– presenta un valor inferior a 0.061 y hay tres items cuyos valores en la segunda columna son inferiores a 0.074 (FNTEA8, 10 y 20). Si se examina la tabla 3 se observa que 8 items de la prueba FNDAT (FNDAT15, 20, 21, 26, 32–34 y 37) presentan un valor para EAcs inferior a 0.066 y hay 10 items con valores inferiores a 0.090 en la segunda columna (FNDAT15, 20, 21, 23, 26, 32–34, 37 y 38).

Conclusiones y Discusión

Al trabajar desde la aproximación factorial se observan resultados consistentes con todos los procedimientos utilizados, que apuntan hacia la ausencia de problemas de sesgo en las pruebas consideradas.

Al trabajar desde la aproximación de la TRI también se observan resultados consistentes en los dos estadísticos del área calculados que apuntan, sin embargo, hacia un problema de sesgo más o menos generalizado en la prueba FNTEA y en algunos items de FNDAT. La falta de consistencia se observa, por tanto, entre los resultados obtenidos desde la aproximación factorial y la TRI.

Esta inconsistencia en los resultados se puede relacionar parcialmente con la naturaleza de los procedimientos utilizados en este trabajo. Por un lado, la aproximación factorial estudia el sesgo de las pruebas considerando éstas de forma global, mientras que la aproximación basada en la TRI supone el cálculo de determinados estadísticos para cada uno de los elementos de las pruebas. Este hecho podría explicar, por ejemplo, que no se detecte sesgo en la prueba FNTEA cuando se utiliza la aproximación factorial, ya que se trataría de una situación en la que podría existir un sesgo generalizado. Por otro lado, la aproximación factorial y la medida del área sin signo serían procedimientos adecuados para detectar sesgo no uniforme mientras que la medida con signo del rea detectaría sesgo uniforme.

Asimismo, un factor de interés que podría arrojar algo de luz a la hora de entender e interpretar los resultados obtenidos sería la consideración del método utilizado para equiparar las estimaciones de los parámetros en las muestras de varones y mujeres. Según Lautenschlager y Park (1988), a juzgar por el pequeño número de estudios que informan del método concreto utilizado en la equiparación, este factor no suele ser considerado importante por muchos investigadores. Sin embargo, como señalan certeramente Cole y Moss (1989), no se puede obviar la posible introducción de error como consecuencia de la equiparación –absolutamente necesaria– de las estimaciones de los parámetros obtenidas en uno y otro grupo. En esta línea, quizás sería interesante examinar si se obtiene el mismo patrón de resultados cuando, en vez de utilizar el método tradicionalmente usado en gran parte de los estudios de la literatura psicométrica –método estándar de la media y la desviación típica–, se utilizan procedimientos iterativos en la equiparación que parecen funcionar bastante bien (Candell y Drasgow, 1988; Candell y Hulin, 1986; Drasgow, 1987; Hulin y Mayer, 1986; Kim y Cohen, 1992; Miller y Oshima, 1992; Park y Lautenschlager, 1990; Segall, 1983).

En cualquier caso, la conclusión inequívoca que sugieren estos resultados es la de la necesidad de un mayor y más profundo estudio de la cuestión. Si bien la literatura evidencia la relativa incapacidad de la aproximación factorial –exploratoria– para detectar sesgo cuando se conoce *a priori* su existencia (Adams y Rowe, 1988), es igualmente cierto que el uso del análisis factorial restringido se ha revelado como un método bastante eficaz para la detección de sesgo (Oort, 1992). Asimismo, resulta obvio que el problema del sesgo no es un problema trivial sino ciertamente complejo. Por ello, es preciso

recoger evidencia múltiple con distintos procedimientos, ya no sólo basados en la TRI sino utilizando otras aproximaciones, como el método de Mantel–Haenszel (Holland y Thayer, 1986, 1988) o la medida omnibus (Johnson, 1989) y trabajar desde la aproximación factorial, pero desde la perspectiva del análisis factorial restringido o confirmatorio. Ahora bien, habida cuenta de que 'en ausencia de criterios teóricos sustantivos, ningún tipo de manipulación estadística de los datos del test puede responder al qué y al porqué del sesgo del item, ni puede facilitar la realización de inferencias válidas acerca de su presencia o ausencia' (Adams y Rowe, 1988, p. 403), resulta evidente que los esfuerzos en la investigación no se deben de dirigir únicamente a cuestiones técnicas relacionadas con la medición del sesgo sino a cuestiones sustantivas relacionadas con la naturaleza del sesgo, qué es y porqué ocurre.

Referencias

Adams, R. J. y Rowe, K. J. (1988): Item Bias. En J.P.Keeves (Ed.), *Educational Research, Methodology and Measurement An International Handbook.* Oxford: Pergamon Press.

Anastasi, A. (1958): *Differential Psychology: Individual and Group Differences in Behavior.* New York: Macmillan.

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.

Candell, G. L. y Hulin, C. L. (1986): Cross-Language and Cross-Cultural Comparisons: Independent Sources of Information about Item Non–Equivalence. *Journal of Cross-Cultural Psychology,* 17, 417-440.

Cole, N. S. y Moss, P. A. (1989): Bias in Test Use. En R. L. Linn (Ed.), *Educational Measurement.* New York: Macmillan.

Drasgow, F. (1987): Study of the Measurement Bias of Two Standardized Psychological Tests. *Journal of Applied Psychology,* 72, 19-29.

Hambleton, R. K. y Rovinelli, R. A. (1973): A FORTRAN IV Program for Generating Examinee Response Data from Logistic Test Models. *Behavioral Science,* 17, 73-74.

Holland, P. W. y Thayer, D. T. (1986): *Differential Item Functioning and the Mantel-Haenszel Procedure* (Technical Report N.° 86-89). Princeton, NJ: Educational Testing Service.

Holland, P. W. y Thayer, D. T. (1988): Differential Item Performance and the Mantel-Haenszel Procedure. En H. Wainer y H. I. Braun (Eds.), *Test Validity.* Hillsdale, NJ: Lawrence Earlbaum Associates, Inc..

Hulin, C. L. y Mayer, L. (1986): Psychometric Equivalence of a Translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology,* 71, 83-94.

Johnson, E. G. (1989): Theoretical Justification of the Omnibus Measure of Differential Item Functioning. En B.J. King, R. Bertrand y F. A. Dupuis, *A World of Differences An International Assessment of Mathematics and Science* (Technical Report).

Keeves, J. P. (1988): Sex Differences in Ability and Achievement. En J. P. Keeves (Ed.), *Educational Research, Methodology and Measurement An International Handbook.* Oxford: Pergamon Press.

Kim, S. H. y Cohen, A. S. (1992): Effects of Linking Methods on Detection of DIF. *Journal of Educational Measurement,* 29, 1, 51-66.

Lautenschlager, G. y Park, D. (1988): IRT Item Bias Detection Procedures: Issues of Model Misspecification, Robustness and Parameter Linking. *Applied Psychological Measurement,* 12, 365-376.

Maccoby, E. E. (Ed.) (1966): *The Development of Sex Differences.* California: Standford University Press.

Miller, M. D. y Oshima, T. C. (1992): Effect of Sample Size, Number of Biased Items and Magnitude of Bias on a Two-Stage Item Bias Estimation Procedure. *Applied Psychological Measurement,* 16, 4, 381-388.

Moss, D. J. (1982): *Towards Equality: Progress by Girls in Mathematics in Australian Secondary Schools.* Australian Council for Educational Research, Hawthorn, Victoria.

Muñiz, J. (1990): *Teoría de respuesta a los items: Un nuevo enfoque en la evolución psicológica y educativa.* Madrid: Pirámide.

Muñiz, J. (1992): *Teoría clásica de los tests.* Madrid: Pirámide.

Oort, F.J. (1992): Using Restricted Factor Analysis to Detect Item Bias. *Methodika,* VI, 150–166.

Park, D. G. y Lautenschlager, G. J. (1990): Improving IRT Item Bias Detection with Iterative Linking and Ability Scale Purification. *Applied Psychological Measurement,* 14, 163–173.

Pine, S. M. (1977): Applications of Item Characteristic Curve Theory to the Problem of Test Bias. En D.J.Weiss (Ed.), *Proceedings of a Symposium presented at the 18th Annual Convention of the Military Testing Association* (Research Report 77-1). Psychometric Methods Program, Department of Psychology, University of Minnessotta.

Raju, N. S. (1988): The Area between Two Item Characteristic Curves. *Psychometrika,* 53, 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.

Rummel, R. J. (1970): *Applied Factor Analysis.* Evanston, IL: Northwestern University Press.

Segall, D. O. (1983): *Test Characteristic Curves, Item Bias, and Transformations to a Common Metric in Item Response Theory: A Methodological Artifact with Serious Consequences and a Simple Solution.* Unpublished manuscript, University of Illinois, Department of Psychology.

Title, C. K. (1988): Test Bias. En J. P.Keeves (Ed.), *Educational Research, Methodology and Measurement An International Handbook.* Oxford: Pergamon Press.

Tyler, L. E. (1956): The Psychology of Human Differences. New York: Appleton-Century-Crofts.

Walker, D. A. (1976): *The IEA Six Subject Survey: An Empirical Study of Education in Twenty-one Countries.* Estocolmo: Almquist y Wiksell.

Warm, T. A. (1978): *A Primer of Item Response Theory.* U.S. Coast Guard Institute Oklahoma City.

Aceptado el 16 del mayo de 1994