Revista de Investigación Educativa, 1997, Vol. 15, n.º 1, págs. 149-168

ERROR TIPO I Y POTENCIA DE LAS PRUEBAS CHI-CUADRADO EN EL ESTUDIO DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

Mª Dolores Hidalgo Montesinos, José A. López Pina y Julio Sánchez Meca[1]
Universidad de Murcia

RESUMEN

En este estudio se compara la tasa de error tipo I y la potencia estadística de tres pruebas chi-cuadrado que se aplican en la evaluación del funcionamiento diferencial de los ítems (FDI): (a) la prueba chi-cuadrado de Scheuneman, (b) la prueba de chi-cuadrado total y (c) la prueba de Mantel-Haenszel. A la vista de los resultados obtenidos, no parece que una prueba sea claramente superior a otra en cuanto a la tasa de error tipo I. Estas parecen ser bastante conservadoras. En cuanto a la potencia, en tamaños muestrales bajos y diferencias mínimas en los índices de dificultad, la prueba de MH fue la más potente, aunque conforme aumenta el tamaño muestral y la magnitud de las diferencias, las tres pruebas convergen, siendo su potencia para detectar el FDI bastante elevada.

Mª Dolores Hidalgo Montesinos. Facultad de Psicología. Campus de Espinardo. Apartado 4071. Universidad de Murcia, 30080-Murcia. Teléfono: (968) 363470. Fax: (968) 364115. E-Mail: mdhidalg@fcu.um.es

1 Parte de este informe fue presentado al III Simposium de Metodología de las Ciencias Sociales y del Comportamiento celebrado en Santiago de Compostela del 12 al 16 de julio de 1993.

ABSTRACT

In this study the Type I error rate and the statistical power of three chi-square tests for detecting and evaluating of differential item functioning (DIF) are compared: (a) Scheunemann chi-square test, (b) Total chi-square test, and (c) Mantel-Haenszel test. The results show that the three procedures have the same type I error rate across all conditions examined. The tests are conservatives. In relation to the statistical power, when sample size was small and differences in difficulty index were minimal, Mantel-Haenszel procedure was the powest. Results indicated that there was close agreement among three DIF detection procedures when the magnitude of differences and the sample size were incremented. In these conditions the statistical power was high.

Los estudios de diferencias individuales han sido objeto de la atención de los investigadores desde los inicios de la Psicología y de la Ciencias de la Educación. En estos estudios se ha tratado de probar si determinadas poblaciones segregadas en función de algunos indicadores (sexo, nivel socioeconómico, raza, etc.) presentan diferencias en algunos constructos de tipo psicológico y/o educativo. Sin embargo, estos estudios han suscitado un amplio rechazo en determinados sectores de la población, que los acusan de estar contaminados por el hecho de emplear instrumentos de medida (tests) que no representan adecuadamente el grado de ejecución de los subgrupos que componen esa población. En concreto, estas acusaciones se centran en afirmar que los tests psicométricos están sesgados, lo cual equivale a decir que, en igualdad de condiciones, los miembros de un grupo obtienen puntuaciones sistemáticamente mayores (o menores) en ese test que los miembros de otro grupo.

Una forma de investigar las razones por las que un grupo es sistemáticamente mejor (o peor) que otro es atender a las características de los ítems. Es decir, puede ser que uno o más ítems no fucionen adecuadamente, de forma que tiende a favorecer a un grupo (grupo de referencia) sobre otro (grupo focal). En este contexto, se dice que el/los ítem/s presenta un fucionamiento diferencial (FDI) en cada uno de los grupos. Estadísticamente, un ítem presenta FDI si «en dos grupos comparables, la probabilidad de obtener una respuesta correcta, en un nivel de habilidad dado, es diferente para cada uno de dichos grupos» (Scheuneman, 1979; Camilli, 1993), o lo que es lo mismo, que en igualdad de condiciones los miembros del grupo de referencia sistemáticamente tienen mayor probabilidad de acertar un ítem que los miembros del grupo focal.

Históricamente el concepto de FDI ha estado unido al de sesgo del ítem, aunque ambos conceptos son distintos. Así, un ítem que presenta FDI estará sesgado si por su contenido favorece a un grupo sobre otro, es decir, es injusto (*unfair*) con uno de los grupos. Cole y Moss (1989) definen el sesgo del ítem como la causa del FDI, es decir, las diferencias observadas en el funcionamiento del ítem son provocadas por algo irrelevante al propósito del test. En términos generales, se distinguen dos tipos de FDI (Mellenbergh, 1982):

- FDI Uniforme o consistente (Camilli y Shepard, 1994), cuando la probabilidad de responder correctamente a un ítem es mayor, a través de todo el rango de habilidad, en un grupo que en otro.
- FDI No Uniforme o inconsistente (Camilli y Shepard, 1994), cuando la probabilidad de responder correctamente a un ítem es mayor en un grupo que en otro, hasta un nivel de habilidad dado. A partir de dicho punto las probabilidades se invierten siendo menor en el primer grupo que en el segundo. En estas situaciones cabe distinguir dos casos: FDI No Uniforme propiamente dicho, cuando se anulan las diferencias en probabilidad entre los dos grupos sometidos a análisis, y FDI Mixto cuando estas diferencias no se anulan.

El FDI es uno de los temas que más investigación está suscitando en el campo de la medida de variables psicológicas y educativas. Con la finalidad de abordar esta problemática se han propuesto distintas aproximaciones que podemos agrupar en dos categorías (Mellenbergh, 1982, 1989; Millsap y Everson, 1993):

- *Métodos de Invarianza Condicional Observada* (ICO) que utilizan las puntuaciones observadas en el test desde la perspectiva del Modelo Clásico. Entre ellos se incluyen procedimientos tales como el estadístico de Scheuneman (1979), el estadístico de Camilli (1979), el estadístico de Mantel-Haenszel (Holland y Thayer, 1988), el procedimiento estandarizado (Dorans y Kulick, 1986; Dorans y Holland, 1993), los modelos loglineales y modelos logit (Marascuilo y Slaughter, 1981; Mellenbergh, 1982), la regresión logística (Bennett, Rock y Kaplan, 1987; Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990), el análisis discriminante logístico (Miller y Spray, 1993) y las variaciones iterativas sobre algunos de ellos (Clauser, Mazor y Hambleton, 1993, 1994; Fidalgo y Paz, 1995a, 1995b; Gómez y Navas, 1995; Holland y Thayer, 1988; Kok, Mellenbergh y Van der Flier, 1985; Navas y Gómez, 1994; Van der Flier, Mellenbergh, Adér y Wijn, 1984).
- *Métodos de Invarianza Condicional No Observada* (ICN). Estos métodos se desenvuelven dentro del marco de la Teoría de la Respuesta al Item, por lo que utilizan las estimaciones de la habilidad, en el modelo logístico apropiado, para evaluar el FDI. Aquí se englobarían procedimientos tales como el estadístico de Lord (1980), las medidas de área (Cohen, Kim y Baker, 1993; Kim y Cohen, 1991; Linn, Levine, Hastings y Wardrop, 1981; Raju, 1988, 1990; Rudner, Getson y Knight, 1980a, 1980b; Shepard, Camilli y Williams, 1984, 1985), los métodos basados en la comparación de modelos a través de estadísticos de razón de verosimilitud (Hoijtink y Molenaar, 1992; Kelderman, 1984, 1989; Kelderman y McReady, 1990; Thissen, Steinberg y Gerrard, 1986; Thissen, Steinberg y Wainer, 1988, 1993), el procedimiento SIBTEST (Shealy y Stout, 1993a, 1993b) y las propuestas iterativas sobre algunos de estos estadísticos (Candell y Drasgow, 1988; Lautenschlager, Flaherty y Park, 1994; Lord, 1980; Miller y Oshima, 1992; Park y Lautenschlager, 1990).

En definitiva, se dispone de un amplio abanico de procedimientos para evaluar el FDI, dependiendo de las características del modelo de medida. Así, si el test ha sido construido bajo el Modelo Clásico, se pueden utilizar, para evaluar el FDI (tanto uniforme como no uniforme), estadísticos tales como el de Scheuneman (1979), el de Camilli (1979) o χ^2 de Pearson y el estadístico de Mantel-Haenszel (Holland y Thayer, 1988). También se pueden utilizar los modelos logit y la regresión logística, aunque estos procedimientos son más apropiados para evaluar el FDI no uniforme (Mellenbergh, 1982; Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990).

Ya que en nuestro país la implantación de la TRI aún se encuentra en sus inicios, los psicólogos y pedagogos tienen que trabajar básicamente con tests construidos bajo el modelo clásico. En esta línea, el objetivo del presente estudio es comparar tres pruebas estadísticas (el estadístico χ^2 de Scheuneman, el estadístico χ^2 Modificado o de Pearson y el estadístico de Mantel-Haenszel) que se adaptan especialmente a las características de estos tests y que han probado tanto en estudios experimentales como en simulación su bondad para evaluar el FDI.

Numerosos trabajos han estudiado el comportamiento del estadístico de Mantel-Haenszel bajo distintas condiciones (Mazor, Clauser y Hambleton, 1991, 1992; Fidalgo y Muñiz, 1995; Kubiak y Colwell, 1990; Raju, Bode y Larsen, 1989; Ryan, 1991), y otros tantos trabajos han comparado la precisión del estadístico de Scheuneman frente al de Camilli (o χ^2 total) u otros procedimientos de evaluación del FDI (Ironson y Subkoviak, 1979; Rudner, Getson y Knight, 1980a, 1980b; Shepard, Camilli y Averill, 1981; Subkoviak, Mack, Ironson y Craig, 1984). No obstante, no se ha realizado un estudio pormenorizado que permita comparar la potencia y tasa de error tipo I de estas tres pruebas. En consecuencia, el objetivo del presente trabajo es evaluar, bajo ciertas condiciones, la potencia y suceptibilidad a cometer errores tipo I de los tres estadísticos: χ^2 de Scheuneman, χ^2 Modificado o de Pearson y el estadístico de Mantel-Haenszel.

Antes de entrar de lleno en el estudio experimental presentaremos brevemente como se organiza la información en estas pruebas, así como sus características esenciales.

I. ORGANIZACIÓN PREVIA DE LA INFORMACIÓN

El cálculo de cualquiera de estos tres estadísticos supone construir una tabla de contingencia para cada ítem en el que se pretende evaluar el FDI. En dicha tabla los sujetos se clasifican según el nivel de habilidad, el grupo de pertenencia (clásicamente dos: referencia y focal) y las posibles respuestas al ítem (en nuestro caso dos: acierto y fallo). A continuación se divide el continuo de habilidad en K intervalos, para después colapsar sobre esta dimensión y establecer K tablas de contingencia bidimensionales (2 x 2), tal y como aparece en la figura 1.

Figura 1
Tabla de contingencia bidimensional en el estudio del FDI

Grupo	Respuesta al ítem j		Total
	Acierto	Fallo	
Referencia	A_k	B_k	n_{Rk}
Focal	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

donde n_{Fk} y n_{Rk} son las frecuencias de sujetos que han contestado el ítem j en el intervalo k, en los grupos focal (F) y de referencia (R); m_{1k} y m_{0k} se corresponden con el número de sujetos que han acertado y que han fallado el ítem j en ambos grupos, respectivamente; A_k y C_k son las frecuencias relativas correspondientes al número de sujetos que han acertado el ítem en cada grupo (referencia y focal); B_k y D_k son las frecuencias relativas correspondientes al número de sujetos que han fallado el ítem en ambos grupos y T_k es el número de sujetos que han contestado al ítem j en el intervalo k.

Aunque existen distintas estrategias para la elección de los intervalos (Donoghue y Allen, 1993), parece que cuando trabajamos con tests de tamaño intermedio o bajo (10 ó 20 ítems), la estrategia más apropiada (Clauser, Mazor y Hambleton, 1994; Donoghue, Holland y Thayer, 1993; Raju, Bode y Larsen, 1989) es dividir el continuo de habilidad en intervalos con la misma amplitud y cuidar que ningún intervalo quede con frecuencias nulas. Este último aspecto se cuidó especialmente, dado que tanto el procedimiento de *MH* como el de χ^2 total (por incluir las frecuencias relativas a la categoría respuesta incorrecta) pueden establecer, en los intervalos que corresponden a sujetos de habilidad elevada, un límite inferior más bajo que el establecido por el procedimiento de Scheuneman.

2. CHI-CUADRADO DE SCHEUNEMAN

En este procedimiento un ítem no presenta FDI "... si la probabilidad de dar una respuesta correcta es la misma para todos los sujetos de igual habilidad, independientemente del grupo de pertenencia" (Scheuneman, 1979). Esta definición supone probar la hipótesis de $p_{Fk}=p_{Rk}$ en los K intervalos, para lo que debemos asumir que la proporción de individuos que responden correctamente al ítem, en un nivel de habilidad dado, es una estimación de la probabilidad de éxito o probabilidad de acertar el ítem en dicho nivel de habilidad. Para probar esta hipótesis, Scheuneman (1979) propone la siguiente prueba estadística:

$$\chi^2_S = \sum_{k=1}^{K} \left[\frac{[A_k - E(A_k)]^2}{E(A_k)} + \frac{[C_k - E(C_k)]^2}{E(C_k)} \right] \quad (1)$$

donde $E(A_k) = n_{Rk}\, m_{1k} / T_k$ y $E(C_k) = n_{Fk}\, m_{1k} / T_k$. Scheuneman (1979) asume que, si la hipótesis nula es cierta, el estadístico sigue una distribución χ^2 con $(K\text{-}1) \times (r\text{-}1)$ grados de libertad, siendo r el número de grupos. Baker (1981) apunta algunos de los problemas de este estadístico:

- Puesto que se considera únicamente la proporción de aciertos, los resultados pueden verse afectados por la presencia de diferencias reales entre grupos (impacto).
- La desigualdad de los tamaños muestrales de los grupos focal y de referencia puede llevarnos a resultados distintos, en función de la equivalencia o no entre estos tamaños muestrales.
- Al considerar únicamente la proporción de aciertos, el estadístico no sigue realmente una distribución χ^2 (Baker, 1981; Marascuilo y Slaughter, 1981).

Sin embargo, Scheuneman (1981) puntualizó, en respuesta a las críticas de Baker (1981), que efectivamente su estadístico no tiene en cuenta la proporción de respuestas incorrectas, y por ello, no sigue una distribución χ^2 si los tamaños muestrales de los grupos a comparar no están equilibrados y/o existen pocas frecuencias en las celdillas de la tabla de contingencia. Indudablemente estos factores afectan directamente a la potencia de la prueba χ^2 (Cohen, 1969). A pesar de estos problemas, Scheuneman concluyó que, utilizándolo con precaución, puede ser una elección acertada, sobre todo por su sencillez de cálculo.

3. CHI-CUADRADO DE PEARSON O TOTAL

El estadístico χ^2 de Pearson es equivalente al propuesto por Camilli (1979, en Ironson, 1982). Se conoce también como estadístico χ^2 modificado o completo (Holland y Thayer, 1988), dado que para el cálculo del mismo se trabaja tanto con los marginales de respuestas correctas como con los marginales de las respuestas incorrectas. Holland y Thayer (1988) proponen la siguiente expresión para el cálculo del χ^2 total:

$$\chi^2_T = \sum_{k=1}^{K} \left[\frac{T_k}{m_{0k}} \frac{[A_k - E(A_k)]^2}{E(A_k)} + \frac{[C_k - E(C_k)]^2}{E(C_k)} \right] \qquad (2)$$

que sigue una distribución χ^2 con K $(r\text{-}1)$ grados de libertad. La ventaja más importante de este estadístico reside en que es preciso conocer el contenido de las celdillas de respuestas incorrectas.

4. ESTADÍSTICO DE MANTEL-HAENSZEL

En la prueba de Mantel-Haenszel (Mantel y Haenszel, 1959) es preciso dividir el continuo de habilidad en K intervalos, generando de este modo tantas tablas de contingencia 2 x 2 como intervalos. Sin embargo, el planteamiento de las hipótesis nula y alternativa es distinto. Así:

$$H_0 : \alpha=1$$
$$H_1 : \alpha \neq 1$$

donde, para un intervalo *k*, el parámetro α es:

$$\alpha = \frac{p_{Rj}\, q_{Fj}}{p_{Fj}\, q_{Rj}} \tag{3}$$

α es un cociente de razones ('odds ratio') o también llamado razón de productos cruzados. Este índice expresa la razón entre la probabilidad de acertar el ítem versus la probabilidad de fallarlo en el grupo focal y la probabilidad de acertar el ítem versus fallarlo en el grupo de referencia. Mantel y Haenszel (1959) proponen como estimador de α:

$$\hat{\alpha} = \frac{A_k D_k / T_k}{C_k B_k / T_k} \tag{4}$$

que puede adoptar valores entre 0 e ∞. Cuando $\hat{\alpha} = 1$ no hay diferencias entre los grupos sometidos a evaluación, por lo que el ítem no presenta FDI. Sin embargo, cuando el valor de $\hat{\alpha} > 1$ nos encontramos ante un ítem que favorece al grupo de referencia sobre el grupo focal. Por último, si $\hat{\alpha} < 1$ el ítem presenta FDI en el grupo focal. El estadístico χ^2 propuesto por Mantel-Haenszel (MH) es:

$$\chi^2_{MH} = \frac{\left(\left|\sum_{k=1}^{K} A_k - \sum_{k=1}^{K} E(A_k)\right| - 0.5\right)^2}{\sum_{k=1}^{K} Var(A_k)} \tag{5}$$

donde

$$Var(A_k) = \frac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)}$$

Si la hipótesis nula es cierta, el estadístico de MH sigue una distribución χ^2 con 1 grado de libertad. Como se puede observar en la ecuación 5 se incluye un factor de corrección por continuidad (0.5) que mejora la aproximación de la prueba χ^2_{MH} a una distribución teórica χ^2_1.

Holland y Thayer (1988) han propuesto una transformación logarítmica de α a una escala simétrica, donde el valor de cero representa la ausencia de FDI; un valor negativo indica que el ítem es más fácil para el grupo de referencia que para el grupo focal, y un valor positivo indica que el ítem es más difícil en el grupo de referencia que en el grupo focal. Esta transformación es de la forma:

$$\Delta_{MH} = -2.35\ \ln(\hat{\alpha}_{MH}) \tag{6}$$

El estadístico α_{MH} puede ser transformado a esta nueva escala utilizando la siguiente expresión:

$$MH - P = p_F - p_R^* \tag{7}$$

donde p_F es la proporción de respuestas correctas observadas en el grupo focal y p_R^* es la proporción de respuestas correctas pronosticadas en el grupo de referencia a partir de la razón de productos cruzados calculada en el estadístico *MH*, que se obtiene a través de la ecuación:

$$p_R^* = [\alpha_{MH} p_F]/[(1 - p_F) + \alpha_{MH} p_F] \quad (8)$$

5. MÉTODO

5.1. Condiciones experimentales

Para realizar este estudio se seleccionaron dos tamaños muestrales (250 y 1000) tanto para el grupo focal como para el grupo de referencia. El tamaño del test fue de 30 ítems y se mantuvo fijo en todas las condiciones experimentales. Para cada uno de los tamaños muestrales se generó una distribución normal de habilidad en el intervalo [-3, +3] con $\mu_\theta = 0$ y σ_θ^2=1, con las rutinas de números aleatorios disponibles en SYSTAT (Wilkinson, 1990). Los 30 ítems del test tuvieron índices de dificultad diferentes y fueron transformados en parámetros de dificultad de la TRI (Hambleton y Swaminathan, 1985; Lord, 1980) con la siguiente ecuación:

$$b_j = \log \frac{(N - A_j)}{A_j} \quad (9)$$

donde N es el tamaño muestral (250 y 1000) y A_j es el número de sujetos que aciertan el ítem j según la dificultad del mismo. Esta transformación fue necesaria para poder obtener las matrices de simulación de respuestas de los 250 y 1000 sujetos en los 30 ítems.

Para el estudio de la tasa de error tipo I se consideraron tres valores de dificultad: 0.2 (ítem difícil), 0.5 (dificultad media) y 0.8 (ítem fácil). Para el estudio de potencia se manipuló la magnitud o cantidad de FDI. Se establecieron cuatro condiciones manipulando las diferencias en dificultad del ítem entre el grupo de referencia y el grupo focal. Estas diferencias fueron de 0.1, 0.2, 0.3 y 0.4. También se controló el índice de dificultad medio de los mismos: dificultad alta (μ_{ID} = 0.20), dificultad media (μ_{ID} = 0.50) y dificultad baja (μ_{ID} = 0.80). Sin embargo, cuando las diferencias en dificultad fueron de 0.3 y 0.4, solamente se tuvieron en cuenta los ítems con dificultad media, ya que en los ítems de dificultad extrema, diferencias tan elevadas como 0.3 y 0.4 provocan que algunas celdillas obtengan frecuencias nulas y, en consecuencia, los estadísticos que deseamos comparar no funcionan adecuadamente. Las cuatro condiciones manipuladas, pues, generaron tres tipos de tests: un test con tres ítems cuyo FDI fue de 0.1, un test con tres ítems cuyo FDI fue de 0.2 y un test con dos ítems cuyas diferencias en dificultad fueron de 0.3 y 0.4, respectivamente.

5.2. Generación de las matrices de datos

Para estudiar la tasa de error tipo I se generó una matriz de respuestas en cada

grupo (referencia y focal) y cada tamaño muestral. Esta matriz fue generada con el programa SIMULA (Hidalgo y López, 1992; v. 1.0) que permite simular respuestas a los ítems con los modelos logísticos de 1-p, 2-p y 3-p, de acuerdo con el procedimiento implementado en Hambleton y Cook (1983).

Ya que la única condición experimental manipulada en este estudio fue la dificultad de los ítems, las matrices fueron simuladas de acuerdo con el modelo de 1-p utilizando la función logística (Birnbaum, 1968; Lord, 1980):

$$P_j(\theta_i) = \frac{1}{1 + exp^{-D(\theta_i - b_j)}} \tag{10}$$

Esta función matemática permite obtener la probabilidad $(P_j(\theta_i))$ de acertar dado un nivel de habilidad. Para ello, el programa SIMULA compara cada uno de los valores de probabilidad con un número pseudo-aleatorio obtenido de una distribución uniforme, cuyos valores se encuentran entre 0 y 1. Si el valor de probabilidad obtenido es menor que el generado de la distribución uniforme, se considera que el sujeto ha fallado el ítem y, por tanto, se asigna un 0 en la casilla correspondiente. Por el contrario, si el valor de probabilidad obtenido es mayor o igual al generado según la rutina pseudo-aleatoria, se considera que el sujeto ha acertado el ítem y, por tanto, se le asigna un 1 a la casilla correspondiente. Con este procedimiento se generaron, para determinar la tasa de error tipo I, 2000 réplicas en cada tamaño muestral: 1000 para el grupo focal y 1000 para el grupo de referencia. En total, se obtuvieron 4000 matrices de datos. Por otro lado, para el estudio de potencia, además de las 2000 réplicas generadas anteriormente para el grupo de referencia, se simularon 1000 réplicas para el grupo focal en cada tamaño muestral y condiciones manipuladas de FDI, lo que hace un total de 6000 matrices de datos simuladas.

5.3. Análisis

Para estudiar la tasa de error tipo I y la potencia estadística se dividió el continuo de habilidad en cuatro intervalos y se procedió a la construcción de las tablas de contingencia correspondientes a cada ítem en cada grupo.

Para obtener la tasa de error tipo I en las tres pruebas se contabilizaron en cada tamaño muestral (250 y 1000) y en los tres ítems sometidos a estudio (niveles de dificultad 0.2, 0.5 y 0.8) el número de rechazos de la hipótesis nula ocurridos en las 1000 réplicas. El criterio para rechazar la(s) hipótesis nula(s) en cada nivel de significación (5% y 1%) fueron los siguientes:

- En la prueba de Scheuneman, a nivel de significación del 5%, se rechazará la hipótesis de no FDI si el valor obtenido supera el especificado en las tablas de χ^2 ($\chi^2_3 = 7.815$); y al nivel de significación del 1%, si $\chi^2_3 = 11.341$.
- En la prueba χ^2 total, la hipótesis de no FDI se rechazará, al 5%, si el valor obtenido es mayor que el especificado en las tablas ($\chi^2_4 = 9.488$), y al 1% si supera el valor de $\chi^2_4 = 13.277$.

- En la prueba de Mantel-Haenszel se rechazará la hipótesis de no FDI, al 5%, si el valor obtenido supera el valor de $\chi^2_1 = 3.841$, y al 1% si supera el valor de $\chi^2_1 = 6.635$.

Para determinar la potencia de estas pruebas se contabilizó, en cada una de las condiciones, el número de veces que se aceptó la hipótesis alternativa, tanto para el nivel de significación del 1% como del 5%. Los criterios de rechazo de la hipótesis nula fueron los mismos que los utilizados en el estudio de la tasa de error tipo I.

6. RESULTADOS

6.1. Tasa de error tipo I

La tabla 1 y las figuras 2 a la 5, presentan las tasas de error tipo I obtenidas en las diferentes pruebas en cada una de las condiciones experimentales, considerando ambos niveles de significación.

En primer lugar, se observa que las tres pruebas (χ^2 de Scheuneman, χ^2 Total y el estadístico de Mantel-Haenszel) controlan bastante bien la tasa de error tipo I en todas las condiciones manipuladas e independientemente del nivel de significación establecido.

Tabla 1
TASAS DE ERROR TIPO I

Prueba	I.D.	5%		1%	
		N = 250	N = 1000	N = 250	N = 1000
Scheuneman	0,2	0,019	0,032	0,000	0,003
	0,5	0,006	0,005	0,000	0,001
	0,8	0,000	0,000	0,000	0,000
Total	0,2	0,020	0,026	0,007	0,004
	0,5	0,041	0,030	0,007	0,012
	0,8	0,015	0,033	0,000	0,006
Mantel-Haenszel	0,2	0,030	0,033	0,005	0,008
	0,5	0,044	0,035	0,004	0,004
	0,8	0,027	0,031	0,006	0,009

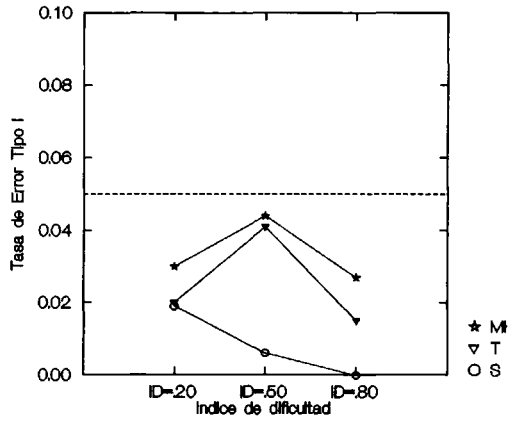

Figura 2
Tasa de error tipo I. N=250 y N.S.=5%

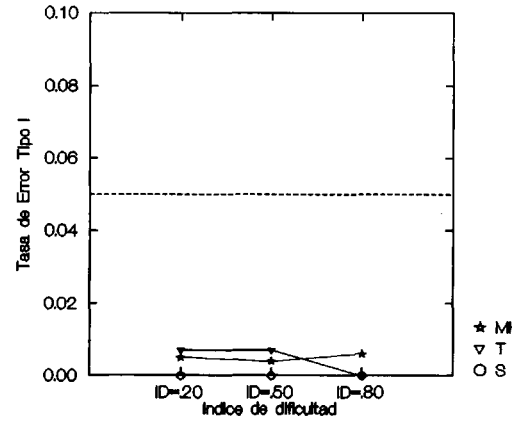

Figura 3
Tasa de error tipo I. N=250 y N.S.=1%

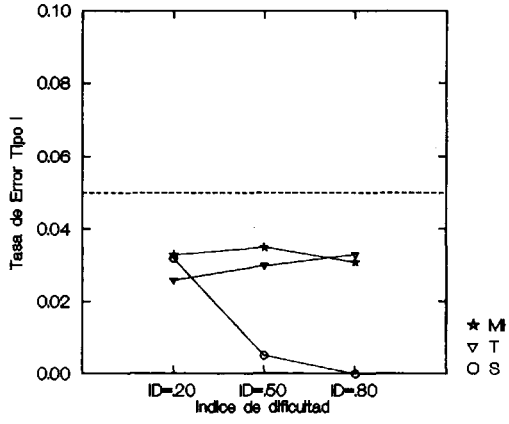

Figura 4
Tasa de error tipo I. N=1000 y N.S.=5%

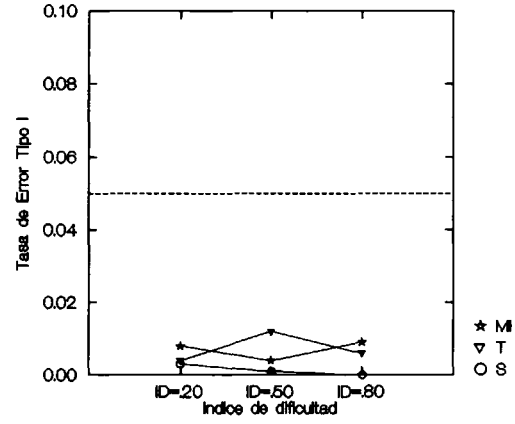

Figura 5
Tasa de error tipo I. N=1000 y N.S.=1%

De la inspección de las figuras 2 y 3 se desprende que cuando la dificultad del ítem fue alta (0.2), las tres pruebas presentan tasas de error tipo I similares. Sin embargo, la situación cambia en los ítems donde la dificultad fue media o fácil. En este caso, y al n.s. del 5%, en tamaños muestrales bajos (N=250), la prueba de Scheuneman obtuvo la tasa de error tipo I más baja, mientras que en las otras dos pruebas incluso se incrementaron. No obstante, si se utiliza el n.s. del 1%, las tres pruebas se igualan (figura 3) y no ofrecen diferencias apreciables en las tres condiciones de dificultad.

Curiosamente el patrón de tasas de error tipo I obtenido cuando el tamaño muestral fue elevado (N = 1000) (figuras 4 y 5) fue el mismo que en la condición anterior, lo

cual evidencia que el tamaño muestral no ha resultado decisivo para mejorar o empeorar la tasa de error tipo I.

6.2. Potencia estadística

En las tablas 2 a la 4 (figuras 6 y 7) se presentan los resultados obtenidos en el estudio de potencia estadística correspondientes a las diferentes pruebas chi-cuadrado en cada una de las condiciones establecidas.

Cuando las diferencias en dificultad fueron de 0.1 (tabla 2), y en tamaños muestrales pequeños (N=250), se aprecia que la prueba con menor potencia estadística fue la de Scheuneman, sobre todo en valores de dificultad medios (0.45-0.55) y bajos (0.75-0.85). En estos casos, y al nivel de significación del 5%, se rechazó la hipótesis de no FDI en 161 veces de las 1000 ocasiones estudiadas cuando la dificultad del ítem fue intermedia y en 88 veces cuando la dificultad del ítem fue baja. Por otro lado, el aumento del tamaño muestral ocasionó una mejora en la potencia de la prueba, donde las tasas de identificaciones correctas llegaron a ser muy altas, con valores de 1 y cercanos a 1.

En la condición de tamaño muestral bajo se puede observar que el estadístico χ^2 total presenta una mayor potencia que la prueba de Scheuneman, aunque las tasas de identificaciones correctas no fueron tan altas como las encontradas con el estadístico de Mantel-Haenszel. El estadístico de Mantel-Haenszel se muestra, bajo estas condiciones, como el más potente para detectar el FDI inducido, sobre todo en valores extremos de dificultad. También se puede observar como en las condiciones de dificultad media ambos estadísticos, χ^2 total y Mantel-Haenszel, tienen menos potencia

Tabla 2
POTENCIA ESTADÍSTICA d = 0,1

Prueba	I.D.	5%		1%	
		N = 250	N = 1000	N = 250	N = 1000
Scheuneman	,15-,25	0,604	1	0,289	0,998
	,45-,55	0,161	0,936	0,045	0,780
	,75-,85	0,088	0,930	0,014	0,710
Total	,15-,25	0,738	1	0,476	0,999
	,45-,55	0,490	0,997	0,252	0,974
	,75-,85	0,730	1	0,495	1
Mantel-Haenszel	,15-,25	0,909	1	0,730	1
	,45-,55	0,669	0,999	0,415	0,997
	,75-,85	0,894	1	0,716	1

Tabla 3
POTENCIA ESTADÍSTICA d = 0,2

Prueba	I.D.	5%		1%	
		N = 250	N = 1000	N = 250	N = 1000
Scheuneman	,10-,30	1	1	0,998	1
	,40-,60	0,940	1	0,741	1
	,70-,90	0,967	1	0,800	1
Total	,10-,30	1	1	0,999	1
	,40-,60	0,997	1	0,984	1
	,70-,90	1	1	1	1
Mantel-Haenszel	,10-,30	1	1	1	1
	,40-,60	1	1	0,998	1
	,70-,90	1	1	1	1

Tabla 4
POTENCIA ESTADÍSTICA d = 0,3 y d = 0,4

Prueba	I.D.	5%		1%	
		N = 250	N = 1000	N = 250	N = 1000
Scheuneman	,35-,65	1	1	1	1
	,30-,70	1	1	1	1
Total	,35-,65	1	1	1	1
	,30-,70	1	1	1	1
Mantel-Haenszel	,35-,65	1	1	1	1
	,30-,70	1	1	1	1

para detectar presencia de FDI. Una posible explicación a este hecho podría ser que las fluctuaciones muestrales de las réplicas obtenidas en el propio proceso de simulación hayan provocado las ligeras diferencias que se aprecian en estas tablas.

Tal y como era de esperar, los estadísticos comparados aumentan su potencia conforme aumenta el tamaño muestral y la magnitud de las diferencias en los índices de dificultad del grupo focal frente al grupo de referencia (ver tablas 3 y 4). Así,

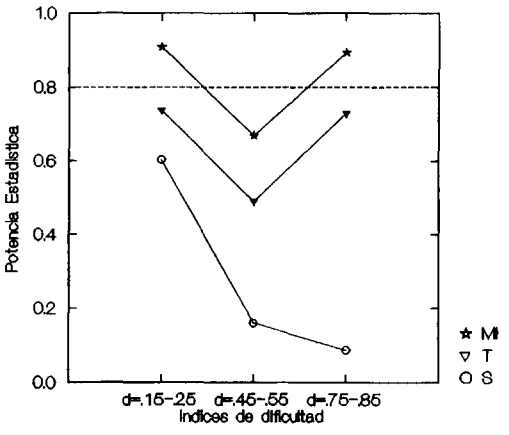

Figura 6
Potencia Estadística. d=0.1, N=250 y N.S.=5%

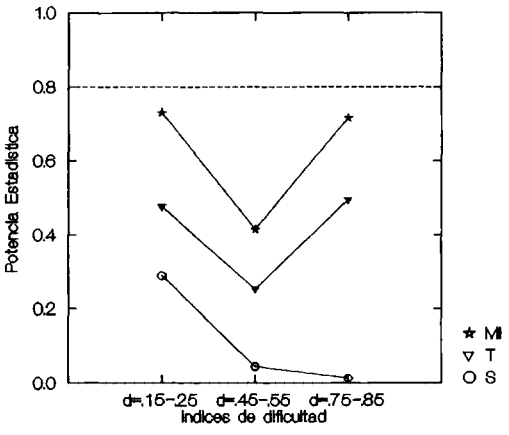

Figura 7
Potencia Estadística. d=0.1, N=250 y N.S.=1%

cuando N = 1000 y la magnitud de FDI fue de 0.2, o superior, todos los estadísticos identificaron correctamente los ítems en los que se había inducido FDI.

7. DISCUSIÓN

A la vista de los resultados obtenidos no parece existir una clara superioridad en cuanto a la tasa de error tipo I de ninguna de las pruebas utilizadas en la evaluación del FDI. Estas parecen ser bastante conservadoras, es decir, ninguna supera la tasa nominal del 5%. No obstante, conviene señalar que, en este nivel de significación y en los ítems con menor dificultad, la prueba de Scheuneman ha obtenido la tasa de error tipo I más baja, frente al estadístico $\chi 2$ total y la prueba de *MH*.

En cuanto a la potencia o tasa de error tipo II, en tamaños muestrales bajos y diferencias mínimas en los índices de dificultad, la prueba de Mantel-Haenszel parece ser la más potente. Pero conforme aumenta el tamaño muestral y la magnitud de las diferencias, no parece existir diferencias entre las tres pruebas, siendo la potencia de las mismas bastante elevada. En este sentido, la mayor potencia del estadístico *MH* está relacionada con la existencia de un sólo grado de libertad, que según Holland y Thayer (1988) hace que el equilibrio entre potencia e identificación de FDI sea mayor que en la prueba de Scheuneman y en la prueba $\chi 2$ total. Estos últimos sólo muestran una gran capacidad para rechazar la hipótesis nula cuando el tamaño muestral es grande, situación que no siempre se da en el estudio del FDI, ya que por regla general, los tamaños muestrales son pequeños y no equilibrados, es decir, el tamaño muestral en el grupo focal es frecuentemente más pequeño que en el grupo de referencia.

Aunque una de las ventajas de estas pruebas reside en que el tamaño muestral necesario para aplicarlas y obtener resultados fiables no necesariamente tiene que ser elevado, como los que se requieren en la aplicación de cualquier procedimiento derivado de la TRI (Kubiak y Colwell, 1990; Mazor, Clauser y Hambleton, 1992), a la vista

de los resultados aconsejamos que se utilicen con tamaños muestrales superiores a 250 sujetos. Mazor, Clauser y Hambleton (1992) aconsejan que bajo presencia de impacto se utilicen con tamaños muestrales superiores a 500 sujetos.

Contrariamente a lo apuntado por otros autores, como Mazor, Clauser y Hambleton (1992), en este estudio, en la condición de menor cantidad de FDI, la capacidad para detectar el FDI fue, en los tres procedimientos, menor cuando los ítems eran de dificultad intermedia que cuando su dificultad fue alta o baja. Sin embargo, en los ítems difíciles (0.15-0.25) las tres pruebas presentaron mayor potencia estadística en comparación con los ítems de dificultad media (0.45-0.55) y de dificultad baja (0.75-0.85). Una posible explicación para estos resultados se puede encontrar en la forma de definir los intervalos de habilidad; es decir, el requisito de que ninguna celdilla de las tablas de contingencia tuviera frecuencias nulas puede haber dado lugar a intervalos de mayor amplitud en los extremos de la escala de habilidad.

Una forma de mejorar la tasa de falsos positivos y ganar en potencia es utilizar un procedimiento iterativo (para el estadístico de *MH*, Holland y Thayer (1988) proponen un procedimiento bietápico), o utilizar el estadístico *MH-modificado* (Fidalgo y Mellenbergh, 1995; Mazor, Clauser y Hambleton, 1994), que además son especialmente útiles en la evaluación del FDI no uniforme (Bock, 1993; Mellenbergh, 1982; Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990).

Por último, pese a que en las condiciones donde tanto el tamaño muestral como el FDI fueron elevados, las tres pruebas presentan la misma potencia (tablas 3 y 4), el estadístico de *MH* resulta más útil en la evaluación del FDI uniforme que χ^2_S y χ^2_T tanto por el acuerdo existente, cuando trabajamos con el modelo logístico de 1-p, entre dicho estadístico y los procedimientos derivados de la TRI (Hambleton y Rogers, 1989; Holland y Thayer, 1998; Thissen, Steinberg y Wainer, 1988; Zwick, 1990) como por la capacidad para detectar presencia mínima de FDI. En cualquier caso resultaría interesante estudiar el comportamiento de estos estadísticos en condiciones de tamaños muestrales no equivalentes y bajo otros criterios de agrupamiento o construcción de intervalos.

REFERENCIAS

Baker, F.B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18*, 59-62.

Bennett, R.E. , Rock, D.A. y Kaplan, B.A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24*, 41-55.

Berk, R.A. (Ed.) (1982). *Handbook of methods for detecting item bias.* Baltimore: Johns Hopkins University Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F.M. Lord y M.R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R.D. (1993). Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. En P.W. Holland y H. Wainer (Eds.) *Differential item functioning* (pp. 115-122). Hillsdale, NJ: LEA.

Camilli, G. (1979). *A critique of the chi-square method for assessing item bias.* Unpublished paper. Laboratory of Educational Research. University of Colorado.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do ítem bias procedures obscure test fairnes issues?. En P.W. Holland y H. Wainer (Eds.) *Differential item functioning* (pp. 397-413). Hillsdale, NJ: LEA.

Camilli, G. y Shepard, L.A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Candell, G.L. y Drasgow, F. (1988). An iterative procedure for linking metrics and assesing item bias in item response theory. *Applied Psychological Measurement, 12,* 253-260.

Clauser, B.E., Mazor, K.M. y Hambleton, R.K. (1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. *Applied Psychological Measurement, 15,* 353-359.

Clauser, B., Mazor, K.M. y Hambleton, R.K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6,* 269-279.

Clauser, B., Mazor, K.M., y Hambleton, R.K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31,* 67-78.

Cohen, A.S., Kim, S.H. y Baker, E. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335-350.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cole, N.S. y Moss, P.A. (1989). Bias in test use. En R.L. Linn (Ed.) *Educational Measurement* (3rd. pp. 201-219). New York: McMillan.

Donoghue, J.R. y Allen, N.L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18,* 131-154.

Donoghue, J.R., Holland, P.W. y Thayer, D.T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. En P.W. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 137- 166). Hillsdale, NJ: LEA.

Dorans, N.J. y Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. En P.W. Holland y H. Wainer (Eds.) *Differential item functioning* (pp. 35-66). Hillsdale, NJ: LEA.

Dorans, N.J. y Kulick, E.M. (1986). Demonstrating the utility of the standardization approach to assesing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Fidalgo, A.M. y Mellenbergh, G.J. (1995). *Evaluación del procedimiento Mantel-Haenszel frente al método logit iterativo en la detección del funcionamiento diferencial de los ítems uniforme y no uniforme.* Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga, Murcia.

Fidalgo, A.M. y Muñiz, J. (1995). *Efectos de la longitud y la dimensionalidad sobre el funcionamiento del procedimiento Mantel-Haenszel.* Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga, Murcia.

Fidalgo, A.M. y Paz, M.D. (1995a). Modelos lineales logarítmicos y funcionamiento diferencial de los ítems. *Anuario de Psicología, 64,* 57-66.

Fidalgo, A.M. y Paz, M.D. (1995b). *Comparación del método logit iterativo frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems.* Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga, Murcia.

Gómez, J. y Navas, M.J. (1995). *Detección del sesgo mediante regresión logística: Purificación paso a paso de la habilidad.* Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga, Murcia.

Hambleton, R.K. y Cook, L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. En D.J. Weiss (Ed.) *New Horizons in testing: Latent trait test theory and computerized adaptative testing* (pp. 31-49). New York: Academic Press.

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,* 313-334.

Hambleton, R.K. y Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hidalgo, M.D. y López Pina, J.A. (1992). *SIMULA 1.0: Un programa para la simulación de vectores de respuesta al ítem.* Laboratorio de Psicometría. Departamento de Psicología Básica y Metodología. Universidad de Murcia.

Hoijtink, H. y Molenaar, I.W. (1992). Testing for DIF in a model with single peaked item characteristic curves: The PARELLA model. *Psychometrika, 57,* 383-397.

Holland, P.W. y Thayer, D.T. (1988). Differential item performance and Mantel-Haenszel procedure. En H. Wainer y H.I. Braun (Eds) *Test validity.* Hillsdale, N.J.: Erlbaum.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. En R.A. Berk (Ed.) *Handbook of methods for detecting item bias* (pp. 117-160). Baltimore: Johns Hopkins University Press.

Ironson, G.H. y Subkoviak, M.J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209-225.

Kelderman, H. (1984). Loglinear Rasch model test. Psychometrika, 49, 223-245.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697.

Kelderman, H. y MacReady, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-327.

Kim, S.H. y Cohen, A.S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15,* 269-278.

Kok, F.G., Mellenbergh, G.J. y van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22,* 295-303.

Kubiak, A.T. y Colwell, W.R. (1990). *Using multiple DIF statistic with the same items appearing in different tests forms.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Lautenschlager, G.J., Flaherty, V.L. y Park, D. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement, 54,* 21-31.

Linn, R.L., Levine, M.V., Hastings, G.N. y Wardrop, J.L. (1981). An investigation of item bias in a test on reading comprehension. *Applied Psychological Measurement, 5,* 159-173.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J.: Erlbaum.

Mantel, N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Marascuilo, L.A. y Slaughter, R.E. (1981). Statistical procedures for identifying posible sources of item bias based on Chi-square statistics. *Journal of Educational Measurement, 18,* 229-248.

Mazor, K.M., Clauser, B.E. y Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443-451.

Mazor, K.M., Clauser, B.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 assesing item bias. *Journal of Educational Statistics, 7,* 105-118.

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

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 method. *Applied Psychological Measurement, 16,* 381-388.

Miller, T.R. y Spray, J.A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

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

Navas, M.J. y Gómez, J. (1994). *Comparison of several bias detection techniques.* Paper presented at the 23rd. International Congress of Applied Psychology, Madrid.

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.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 492-502.

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

Raju, N.S., Bode, R.K. y Larsen, V.S. (1989). An empirical assesment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2,* 1-13.

Rogers, H.J. y Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

Rudner, L.M., Getson, P.R. y Knight, D.L. (1980a). A montecarlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1-10.

Rudner, L.M., Getson, P.R. y Knight, D.L. (1980b). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213-233.

Ryan, K.E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement, 28,* 325-337.

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

Scheuneman, J.D. (1981). A response to Baker's criticism. *Journal of Educational Measurement, 18,* 63-66.

Shealy, R.T. y Stout, W.F. (1993a). An item response theory model for test bias and differential test functioning. En P.W. Holland y H. Wainer (Eds.) *Differential item functioning,* (pp. 197-239). Hillsdale, NJ: LEA.

Shealy, R.T. y Stout, W.F. (1993b). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shepard, L.A., Camilli, G. y Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317-375.

Shepard, L.A., Camilli, G. y Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93-128.

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

Subkoviak, M.J., Mack, J.S., Ironson, G.H. y Craig, R. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21,* 49-58.

Swaminathan, H. y Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Thissen, D., Steinberg, L. y Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118-128.

Thissen, D., Steinberg, L. y Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. En H. Wainer y H.I. Braun (Eds.) *Test validity.* Hillsdale, N.J.: Erlbaum.

Thissen, D., Steinberg, L. y Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. En P.W. Holland y H. Wainer (Eds.) *Differential item functioning* (pp. 67-113). Hillsdale, NJ: LEA.

Van der Flier, H., Mellenbergh, G.J., Adér, H.J. y Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21,* 131-145.

Wainer, H. y Braun, H.I. (Eds) (1988). *Test validity.* Hillsdale, N.J.: Erlbaum.

Wilkinson, L. (1990). SYSTAT: *The system for Statistics* (V. 5.0). Evanston, IL: Systat Corporation.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185-197.