Avances en Medición, 5, 27–46
2007

DETECCION DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ITEMS EN EL MARCO DE LA TEORIA DE RESPUESTA AL ÍTEM[1]

Aura Nidia Herrera[*]
Universidad Nacional de Colombia, Colombia

Juana Gómez[**]
Universidad de Barcelona, España

José Muñiz[***]
Universidad de Oviedo, España

Resumen

Este trabajo es una revisión de las principales propuestas metodológicas basadas en la Teoría de Respuesta al Ítem (TRI) para la detección del Funcionamiento Diferencial de los Ítems (DIF, iniciales de Differential Item Functioning) de calificación dicótoma con modelos TRI unidimensionales y con dos grupos de comparación. En un primer momento se presentan los desarrollos de procedimientos tendientes a garantizar que la comparación de las estimaciones de los parámetros de los modelos TRI, obtenidas a partir de los dos grupos, sea adecuada y precisa; se presentan entonces los procedimientos de equiparación de puntuaciones y de purificación del criterio de igualación de los grupos. Posteriormente se revisan las principales propuestas de identificación de ítems con DIF, mediante comparación de las estimaciones para los grupos; aquí se incluyen los procedimientos de comparación de modelos, comparación de parámetros y las medidas del área de discrepancia entre las Curvas Características de los Ítems (CCI). A partir de esta revisión se tiene una panorámica general del estado actual de los temas tratados, de los aspectos sobre los cuales no hay resultados concluyentes y de las posibilidades de desarrollo en el futuro cercano.

Palabras clave: Teoría de respuesta al ítem, Funcionamiento diferencial de los ítems, medidas de área entre CCI, Ji cuadrado de Lord, TRI.

Abstract

This paper reviews the main methodological approaches based on item response theory (IRT) for detecting differential item functioning (DIF), focusing on dichotomous items with one-dimensional IRT models and two comparison groups. The authors begin by describing the development of procedures aimed at ensuring that the comparison of parameter estimates from IRT models, obtained from the two groups, is adequate and precise. Then they turn to the procedures for comparing scores and purifying the matching criterion of the groups. This is followed by a review of the main approaches for identifying items with DIF by comparing the estimates for each group; this includes procedures for comparing models, for comparing parameters and for measuring the area between the item characteristic curves (ICCs). Thus, providing a general overview of current opinion on the issues addressed, and considering those aspects for which conclusive results are not available yet, and the possibilities for progress in the near future.

Key words: Item response theory, differential item functioning, area between the ICCs, Lord's Chi-squared, IRT, DIF.

---

[1] Este trabajo constituye un producto de investigación del grupo "Métodos e instrumentos para investigación en salud" y ha sido posible gracias a la financiación del Ministerio de Ciencia y Tecnología de España y los fondos FEDER (proyectos SEJ2005-08924 y SEJ2005-09144- C02-02). Los autores agradecen al profesor Agustín Tristán y a un lector anónimo por sus valiosos comentarios al manuscrito.

[*] Laboratorio de Psicometría. Departamento de Psicología. Universidad Nacional de Colombia, Ciudad Universitaria. Bogotá-Colombia. E-mail: anherrerar@unal.edu.co

[**] Departamento de Metodología de Ciencias del Comportamiento. Facultad de Psicología. Universidad de Barcelona. Passeig Vall d'Hebrón, 171. 08035 Barcelona-España. E-mail: juanagomez@ub.edu

[***] Facultad de Psicología. Universidad de Oviedo. Plaza Feijoó, s/n. 33003 Oviedo-España. E-mail: jmuniz@uniovi.es

Los orígenes de la Teoría de Respuesta al Ítem (TRI) pueden encontrarse en los trabajos de Richardson (1936), Lawley (1943), Guttman (1944), Tucker (1946), Lazarsfeld (1950) y Lord (1952), quienes formularon algunos de sus fundamentos; sin embargo, el mérito del desarrollo de los modelos más utilizados hoy correspondió a Rasch (1960) y Birnbaum (1968). Después de la publicación de estos últimos trabajos y hasta los primeros años del presente siglo se ha observado un importante número de publicaciones referidas a tópicos específicos de los modelos de Rasch y Birnbaum y a la generación de otros aplicables a diversas condiciones de medida (Gómez-Benito & Hidalgo-Montesinos, 2003). Un tema que suscitó interés en las décadas subsiguientes fue la generación de algoritmos de estimación de los parámetros de los modelos y la precisión de los mismos, algunos de los trabajos importantes al respecto son los de Bock & Lieberman (1970), Lord (1980), Bock & Aitkin (1981), Mislevy (1986), Mislevy & Bock (1990) y más recientemente Fox (2003). Por otra parte, surgió el desarrollo de rutinas o programas de computador para la estimación de los parámetros y calibración de ítems, algunos de ellos son el LOGIST de Wood, Wingersky & Lord (1976), el BICAL de Wright & Mead (1976), el BILOG de Mislevy & Bock (1990), el DRAWICC de DeMars (2000), el EPDIRM de Hanson (2001) y el WINSTEPS de Linacre (2003), entre muchos otros. Finalmente, un gran número de trabajos se han dedicado a explorar posibles aplicaciones de los modelos TRI en diferentes contextos o al desarrollo de otros modelos para responder a una diversidad de condiciones de medida. Algunas de las revisiones sobre este tema son las de Lord, (1980), Hulin, Drasgow & Parsons (1983), Muñiz & Hambleton (1992) y Santor & Ramsay (1998).

Desde la perspectiva de la TRI, si una prueba es esencialmente unidimensional[2] y dada una magnitud de atributo (θ) la probabilidad de acertar en un ítem es independiente de la probabilidad de acertar en los demás[3], entonces la probabilidad condicional de acierto en el mismo para un nivel de magnitud de rasgo latente ($P(U_i \mid \theta)$, que suele notarse sencillamente como $P_i(\theta)$) es una función monotónica creciente que puede expresarse como un modelo logístico de la forma

$$P_i(\theta) = f_i + \frac{e^{dx}}{1+e^{dx}},$$

donde d es una constante cuyo valor se asigna dependiendo del tipo de modelo que se desee ajustar, el parámetro aditivo *fi* depende de una característica del ítem y x es una expresión que involucra los parámetros de los ítems que se consideren en cada caso. Históricamente se han trabajado dos tipos de modelos, el primero de ellos propuesto por Lord (1952), se basa en la distribución normal acumulada y el segundo tipo, desarrollado gracias a los trabajos de Rasch (1960) y Birnbaum (1968), se basa en la ya conocida función logística. Actualmente los modelos de ojiva normal tienen poco uso debido a que los logísticos resultan más sencillos y económicos; sin embargo *d=1,7* permite una aproximación bastante precisa del modelo logístico a la función

---

[2] El supuesto de dimensionalidad se refiere a la definición completa de los rasgos latentes –espacio latente– que evalúa la prueba; sin embargo, los modelos de la TRI que han tenido mayor desarrollo son los que miden sólo un rango latente, conocidos como modelos unidimensionales o esencialmente unidimensionales. Este trabajo se limita a este tipo de modelos, en Cuesta (1996), Harvey & (1999), Cuesta & Muñiz (1994, 1995) y Drasgow & Parsons (1993) se pueden encontrar explicaciones y discusión detenidas sobre el tema.

[3] Este supuesto se conoce como independencia local y estadísticamente se expresa como el producto de probabilidades: para un nivel de magnitud de atributo la probabilidad de puntuar en un conjunto de ítems es igual al producto de las probabilidades de puntuar en cada uno de ellos. En Crocker & Algina (1986), Lord, (1980) y Lord & Novick (1966) se encuentra un tratamiento más detenido de este supuesto y su relación con la unidimensionalidad.

normal acumulada. De otra parte, hasta el momento se han considerado predominantemente tres parámetros de los ítems: dificultad, discriminación y pseudoazar – también conocido como pseudoadivinación o adivinación sistemática–, notados para un ítem *i*, como b_i, a_i y c_i, respectivamente. Si se considera que todos son relevantes para explicar la relación entre θ y $P_i(\theta)$, se tendrá un modelo logístico de tres parámetros (3P) de la forma $P_i(\theta) = c_i + (1 - c_i)\frac{e^{dx}}{1+e^{dx}}$ con $x = a_i(\theta - b_i)$. Si se supone que el ítem no puede responderse acertadamente por azar o que esta probabilidad es despreciable, entonces se ajusta un modelo dos parámetros (2P) haciendo $c_i = 0$. Finalmente, en el modelo de un parámetro (1P) se supone además que la discriminación es constante para todos los ítems, por ejemplo $a_i = 1$, y entonces la expresión queda sencillamente como $P_i(\theta) = \frac{e^{d(\theta - b_i)}}{1 + e^{d(\theta - b_i)}}$. Esta expresión se conoce como la función característica del ítem (FCI) y su representación gráfica es la curva característica del ítem (CCI).

El parámetro de dificultad se define en la misma escala de la magnitud de atributo, es el valor de θ en el punto de máxima pendiente de la CCI y define la posición de la CCI sobre la escala de θ. El parámetro de discriminación se define como la capacidad del ítem para distinguir entre examinados con altas y bajas magnitudes de atributo; puede asimilarse al nivel de inclinación de la curva y es proporcional a la pendiente de la recta tangente a la CCI en su punto de mayor inclinación. Finalmente, el tercer parámetro se define como la probabilidad de puntuar en un ítem por azar, corresponde a la probabilidad de acierto cuando θ es mínimo; sin embargo, dado que θ puede tomar valores entre $-\infty$ y $+\infty$, *c* corresponde a la asíntota de la CCI cuando θ tiende a $-\infty$ pero este valor asintótico no es observable en las CCI puesto que generalmente se fija un valor mínimo de *-3* o *-4* para la escala de θ; así, el parámetro de pseudoazar puede verse como la mínima probabilidad de acertar.

Que θ pueda tomar valores entre $-\infty$ y $+\infty$ significa que, en teoría, los modelos TRI pueden estar localizados en cualquier intervalo de valores sobre la recta real sin que esto afecte la relación entre la magnitud de atributo y la probabilidad de acertar al ítem. En la práctica esto implica que una vez estimados, los valores de los parámetros del modelo pueden transformarse para lograr expresiones en una escala convencional que facilite, por ejemplo, la comparación entre pruebas o entre grupos. Obviamente cambiar la escala que se utilizó originalmente para las estimaciones, implica transformar también los valores de las estimaciones de los parámetros de los ítems. Un tipo de transformación lineal frecuentemente utilizado es de la forma $\theta' = m\theta + k$ con $m > 0$; en este caso los parámetros de dificultad y discriminación quedan como $b_i' = mb_i + k$ y $a_i' = a/m$, respectivamente, mientras que el valor de c_i no cambia. Pero esta transformación lineal no es más que un ejemplo de los muchos tipos de transformaciones de la escala, que podrían utilizarse; tanto Muñiz (1997) como Hambleton, Swaminathan & Rogers (1991) explican los tipos de transformaciones más utilizadas. Ya que este tema es de vital importancia en el uso de técnicas basadas en la TRI para detectar ítems sesgados, más adelante se tratará con algo más de detalle.

Una de las aplicaciones prácticas de la TRI ha sido el desarrollo de procedimientos para detectar el posible sesgo en los instrumentos de medida, tema que generó tanta polémica como interés político, social y académico en las últimas décadas del pasado siglo (Gómez-Benito & Hidalgo-Montesinos, 2003). Aunque los trabajos de Stern (1914) y Binet & Simon (1916) habían hecho notar algunas diferencias en el desempeño en las pruebas entre personas pertenecientes a grupos culturales diferentes, la literatura científica reconoce en Eelles, Havighurst, Herrick & Tyler (1951) y Jensen (1969) las publicaciones pioneras y trascendentales en los desarrollos posteriores sobre el tema. Muñiz (1997) hace notar cómo las publicaciones psicométricas especializadas de las décadas de los cincuenta y los sesenta del pasado siglo, y la edición de 1966 de los *Standards for Educational and Psychological Test and Manuals* ignoran por completo el tema, y sólo a partir de los 70 la comunidad psicométrica se apropia de la discusión que se había mantenido en las esferas legal, política, social y de la teoría psicológica.

Las tres últimas décadas del pasado siglo y los primeros años del naciente han sido testigos de un gran interés por construir un marco conceptual sólido y desprovisto de connotaciones éticas, sociales o políticas, para los estudios de sesgo en los instrumentos de medida, y por generar estrategias metodológicas eficientes para detectarlos. Dentro de la primera categoría de trabajos, algunos de los de mayor trascendencia son los de Jensen (1980), Reynolds (1982), McCauley & Mendoza (1985), Holland & Thayer (1988), Kok (1988), Drasgow (1987), Shealy & Stout (1989), Ackerman (1992), Holland & Wainer (1993) y Borsboom, Mellenbergh & van Heerden (2002). Por otra parte, dentro de las propuestas metodológicas más importantes para detectar los instrumentos posiblemente sesgados cabe citar las de Angoff (1972), Angoff & Ford (1973), Lord, (1980), Mellenbergh (1982), Dorans & Kulick (1986), Bennet, Rock & Kaplan (1987), Thissen, Steinberg & Wainer (1988, 1993), Holland & Thayer (1988), Raju (1988, 1990), Dorans (1989), Shealy & Stout (1989, 1993a, 1993b), Swaminathan & Rogers (1990), Kim & Cohen (1991) y Cohen, Kim & Baker (1993).

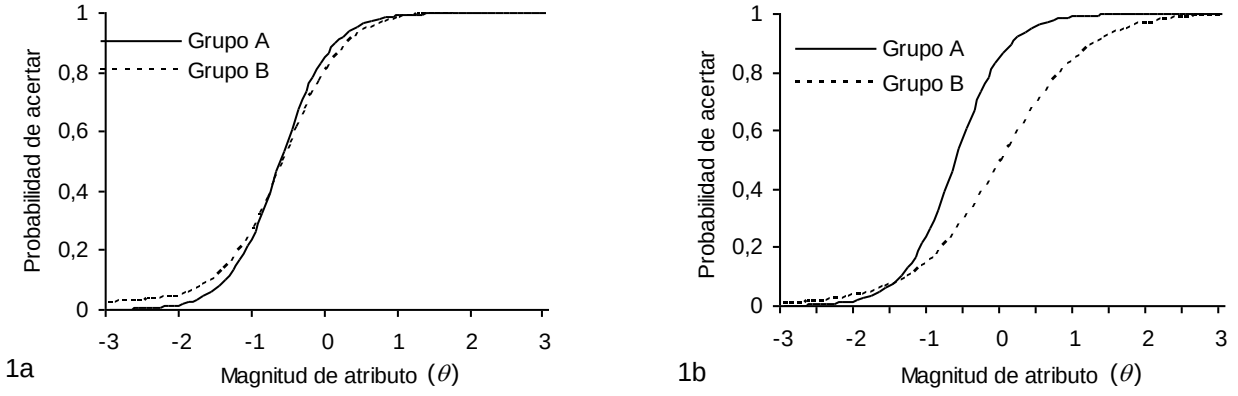

Figura 1. CCI de dos ítems para dos grupos diferentes. 1a: Ítem sin DIF. 1b: ítem con DIF

Hoy, la expresión DIF, iniciales de *Differential Item Functioning*, es muy popular en la jerga psicométrica para referirse a la evidencia empírica, soportada estadísticamente, de que un ítem funciona de manera diferente para grupos diferentes. Brevemente, un ítem tiene DIF si la probabilidad de puntuar en el mismo cambia no solamente en función de la magnitud de atributo sino del grupo de pertenencia, es decir, tiene diferente comportamiento entre personas con igual magnitud del atributo pero que pertenecen a grupos diferentes. La figura 1 representa las CCI de dos ítems para dos grupos. La probabilidad de acertar en el ítem de la izquierda (1a) es muy

similar para individuos con la misma magnitud de atributo aún cuando pertenezcan a grupos diferentes, se trata de un ítem no sesgado. Por el contrario, las CCI del ítem de la figura 1b cambian en función del grupo haciendo que la probabilidad de acierto sea diferente para individuos con la misma magnitud de atributo, se trata de un ítem que presenta DIF.

Los métodos que se basan en la TRI para la detección de DIF se sustentan justamente en la comparación de las CCI de un ítem cuando se ajustan modelos separadamente para los dos grupos de interés. Brevemente, estos procedimientos ajustan modelos TRI para los dos grupos independientemente, expresan los parámetros en la misma métrica y comparan los dos modelos. La idea de fondo es que el DIF se manifiesta mediante diferencias en las CCI del ítem cuando se calcula para los dos grupos separadamente; así, si un ítem no tiene DIF, las CCI de los dos grupos deben coincidir salvo por pequeñas variaciones atribuibles al azar como se mostró en la figura 1a. Pero el proceso de detección de ítems con DIF desde la perspectiva de la TRI exige abordar previamente dos requerimientos metodológicos que garanticen la comparabilidad de las estimaciones obtenidas para los dos grupos: equiparación y purificación de la escala.

Requerimientos previos

Se mencionó antes que las estimaciones de la magnitud del atributo de los individuos, θ, y por ende el parámetro de dificultad de los ítems, *b*, pueden expresarse en cualquier rango de valores entre $-\infty$ y $+\infty$; esto implica que la comparación entre dos estimaciones del mismo parámetro solamente será adecuada si se puede garantizar que estén expresadas en la misma escala, es decir, si se implementa algún procedimiento de equiparación. De otra parte, la identificación de DIF implica comparar individuos con la misma magnitud de atributo y en la práctica solamente se dispone de las respuestas observadas de los individuos a un grupo de ítems, incluyendo los posiblemente sesgados; en consecuencia, otro aspecto debe controlarse en el proceso es el posible efecto de los ítems sesgados en las estimaciones de θ, conocidos como procedimientos de purificación de la escala. A continuación se tratarán separadamente los dos temas.

Equiparación de puntuaciones

Angoff (1984) y Lord (1980), citados frecuentemente como los pioneros en el tratamiento del tema, entienden la equiparación como un proceso que implica expresar el sistema de unidades de una prueba en el de otra, de manera que sus resultados sean comparables o equivalentes; el problema consiste entonces en hallar una escala o métrica común para dos o más medidas de un mismo atributo de manera que se puedan comparar resultados de individuos o de instrumentos diferentes (Kolen, 2004). En teoría, la propiedad de invarianza de los parámetros de la TRI implicaría que en este marco, el tema carece de toda importancia ya que, conocidos los parámetros de los ítems, la estimación de la magnitud de atributo es independiente del grupo de ítems incluidos en la prueba; y a su vez, conocida la magnitud de atributo de los individuos, la estimación de los parámetros de los ítems es independiente del grupo de examinados. Desde esta perspectiva, dada la indeterminación de la escala de θ, Hambleton et al.,(1991) sugieren que en el marco de la TRI no se hable de equiparación (*equating*) sino de 'escalamiento' (*scaling*) para referirse a la necesidad de elegir una escala común para las estimaciones de los parámetros tanto de los ítems como de los individuos, cuando éstas se basan en pruebas o grupos diferentes.

Háblese de equiparación o de escalamiento, en la práctica jamás se tendrán los mismos valores de los parámetros cuando éstos se estiman a partir de las respuestas de dos grupos diferentes, y en la detección de DIF resulta de gran importancia garantizar la comparabilidad entre tales estimaciones, lo que implica tenerlos expresados en la misma métrica. En el marco de la TRI el proceso de equiparación puede entenderse como una transformación lineal de los parámetros, del tipo que se presentó anteriormente: $\theta' = m\theta + k$, $a_i' = a_i / m$, $b_i' = mb_i + k$ y $c_i' = c_i$, sin alterar la probabilidad de acierto al ítem. Así, el proceso consiste en encontrar los valores de las constantes *m* y *k* que permitan expresar con mayor precisión unas estimaciones en la métrica de otras. La estrategia más utilizada es la denominada "prueba de anclaje" que consiste en incluir en las dos aplicaciones un grupo de ítems comunes que constituyen la base para la equiparación puesto que se dispone de dos estimaciones de sus parámetros. Este procedimiento también es aplicable cuando se incluyen 'individuos de anclaje' y se dispone de dos estimaciones de su magnitud de atributo; la siguiente exposición se enmarca en el primer caso, no sin mencionar que las transformaciones sobre *b*, son aplicables a θ.

Disponiendo de dos estimaciones de los parámetros de un mismo grupo de ítems, puede calcularse el valor de las constantes de equiparación optando por uno de los métodos tradicionales[4] o los basados en la TRI. Dentro de estos últimos se encuentran el método de la función de respuesta al ítem y el χ^2 mínimo. El primero, propuesto originalmente por Haebara (1980 citado en Navas Ara, 1996) y Stocking & Lord (1983), deriva los valores de las constantes de equiparación minimizando las diferencias entre las puntuaciones verdaderas de los individuos, en las dos estimaciones. Si τ_{jX} y τ_{jY} son las puntuaciones verdaderas del individuo *j* en los ítems o subpruebas de anclaje *X* y *Y*, respectivamente; las constantes de equiparación, *m* y *k*, serán los valores que minimicen la función $F = \frac{1}{n}\sum_{j=1}^{n}\left(\tau_{jX} - \tau_{jY}\right)^2$. Siguiendo el mismo razonamiento, Divgi (1985) propuso el método conocido como χ^2 mínimo ya que minimiza la función de las diferencias entre las estimaciones de los parámetros en las dos aplicaciones, función que coincide con el estadístico χ^2 de Lord para la detección de DIF y que se presentará con algo más de detalle en la siguiente sección. Más recientemente Ogasawara (2001) propuso una estimación por mínimos cuadrados que no ha sido muy evaluada hasta el momento.

Hoy se dispone de una buena cantidad de trabajos que evalúan la eficiencia relativa de cada uno de estos métodos o las condiciones bajo las cuales resultan más recomendables; entre ellos se pueden citar Candell & Drasgow (1988), Lautenschlager & Park (1988), Baker & Al-Karni (1991), Kim & Cohen (1992), Millsap & Everson (1993), Baker (1996), Kaskowitz & De Ayala (2001), Hidalgo Montesinos & López Pina (2002), entre muchos otros. Sin embargo, los resultados no son concluyentes en favor de uno u otro procedimiento. Algunas investigaciones (Stocking & Lord, 1983; Kim & Cohen, 1992, 1994) sugieren que el método de función de respuesta al ítem puede

[4] Los métodos no basados en la TRI son el de regresión y los de media–desviación. Para una explicación detallada sobre los mismos puede consultarse Linn, Levine, Hastings & Wardrop (1981), Stocking & Lord (1983) o Navas Ara (1996)

resultar más preciso comparativamente y Kaskowitz & De Ayala (2001) encontraron que este procedimiento tal como lo implementa el EQUATE de Baker (1995), es relativamente robusto a la magnitud de error de estimación de los parámetros. Por el contrario, Candell & Drasgow (1988) encontraron que un procedimiento no basado en la TRI, más fácil y económico, mostraba mejores resultados mientras que para Park & Lautenschlager (1990) el mejor método de equiparación, considerando precisión y costos, es el χ^2 mínimo. Baker & Al-Karni (1991) por su parte, no encontraron diferencias entre métodos no basados en la TRI y los que comparan las CCI.

Purificación de la escala

El otro aspecto importante para garantizar una adecuada precisión en la identificación de los ítems con DIF hace referencia al posible 'efecto contaminante' de estos ítems en la estimación de la magnitud de atributo como criterio de igualación de los grupos, y en el cálculo de las constantes de equiparación. Para controlar este efecto se han propuesto procedimientos de 'purificación' que buscan, en varias etapas o iterativamente, identificar los ítems con DIF y excluirlos para efectos de la estimación de θ o del cálculo de *m* y *k*. La propuesta pionera en este sentido fue la publicada en el texto de Lord, sugerida por Marco (1977, en Lord, 1980), que implica comparar las CCI de todos los ítems para identificar los que presentan DIF, estimar θ a partir de los ítems que no presentan DIF uniendo los dos grupos, estimar los parámetros de los ítems utilizando como valores fijos estas estimaciones de θ, y comparar nuevamente las CCI con las estimaciones 'purificadas' de los parámetros de los ítems. Posteriormente Segal (1983, citado en Candell & Drasgow, 1988) propuso un procedimiento iterativo que va ajustando los valores de *m* y *k* obviando la re-estimación de θ; éste consiste en realizar una equiparación con base en estimaciones iniciales de los parámetros para los dos grupos separadamente, identificar los ítems con DIF, hacer una nueva equiparación excluyéndolos, y recalcular los índices de DIF; estos dos últimos pasos se repiten hasta que en dos iteraciones consecutivas se identifiquen los mismos ítems con DIF. Drasgow, (1987) utilizó este procedimiento con el método de la función de respuesta al ítem para el cálculo de las constantes de equiparación y un año más tarde Candell & Drasgow (1988) demostraron su efectividad utilizando además métodos no basados en la TRI para calcular *m* y *k*.

Por su parte, Park & Lautenschlager (1990) propusieron una combinación de las dos propuestas anteriores; ésta se inicia con el ajuste de las constantes de equiparación siguiendo el procedimiento iterativo de Segal (1983) y el método de cálculo del χ^2 mínimo, una vez se logra el criterio de convergencia se re-estima θ para los dos grupos separadamente a partir de los ítems insesgados, entonces se re-estiman los parámetros de los ítems utilizando como valores fijos estas estimaciones de θ y finalmente, se identifican los ítems con DIF. De acuerdo con Lautenschlager, Flaherty & Park (1994) este procedimiento arroja resultados más satisfactorios que los dos anteriores; aunque Miller & Oshima (1992) y Hidalgo Montesinos & López Pina (2002) lo consideran demasiado costoso en términos de cómputo. En consecuencia Miller & Oshima (1992) propusieron una modificación que consiste en un procedimiento de dos etapas que obvia el proceso iterativo inicial para calcular las constantes de equiparación pero calcula dichas constantes y estima los parámetros de los ítems dos veces. Hidalgo Montesinos & López Pina

(2002) sintetizaron esta propuesta en un procedimiento de dos etapas con una sola estimación de los parámetros de los ítems. En la primera etapa se estiman estos parámetros para los dos grupos separadamente, se equiparan las métricas y se identifican y eliminan los ítems con DIF; en la segunda etapa se realiza la equiparación con los ítems libres de DIF y se obtiene el estadístico para detectar DIF para todos los ítems. Mediante un estudio con datos simulados, los autores mostraron que este procedimiento arroja resultados satisfactorios utilizando el χ^2 de Lord (1980) y las medidas de áreas de Raju (1988, 1990) para la detección del DIF.

Sea cual sea el procedimiento de purificación elegido, una estrategia frecuentemente utilizada en los estudios de DIF es calcular las constantes de equiparación tomando como subprueba de anclaje todos los ítems de la prueba con excepción del que está siendo estudiado; sin embargo recientemente Wang & Yeh (2003) y Wang (2004) mostraron que este procedimiento arroja resultados satisfactorios solamente cuando ningún ítem de la prueba presenta DIF o si el ítem analizado es el único que lo presenta o si hay varios ítems con DIF pero unos favorecen a un grupo y otros al otro grupo de manera que se presente cancelación de los efectos. Dado que estas condiciones son difíciles de satisfacer en la práctica o, si se cumplen, hacen innecesario el estudio, resulta más recomendable usar unos ítems determinados como subprueba de anclaje para el análisis de todos los demás. Esto implica decidir el mecanismo apropiado para la selección de dichos ítems y el número de ítems necesarios para obtener un cálculo preciso de las constantes de equiparación. La mejor estrategia para la selección de los ítems de anclaje sería la identificación de algunos previamente calibrados y con evidencia de que no tienen DIF, sin embargo, esto sólo es posible si se dispone de bancos suficientemente amplios y analizados, condición que no se cumple en muchas situaciones reales (Fidalgo, 1996; Navas-Ara, 1996). Algunos autores (Thissen et al., 1988; Wang & Yeh, 2003) han sugerido utilizar un procedimiento menos costoso que los basados en la TRI[5], para identificar algunos ítems insesgados que puedan utilizarse como subprueba de anclaje para la equiparación, y Wang, (2004) propuso un procedimiento iterativo para la selección de los ítems de anclaje, sin embargo, este último no ha sido aún evaluado. De otra parte, dos trabajos recientes se han ocupado del examinar el efecto del número de ítems de anclaje sobre la precisión de la equiparación: Según Kaskowitz & De Ayala (2001) 5 ítems es muy poco y recomiendan utilizar 15 o 25 para obtener resultados satisfactorios, sin embargo, Wang & Yeh (2003) sugieren que 4 ítems, si realmente están libres de DIF, pueden ser suficientes para obtener una precisión adecuada.

Identificación de los ítems con DIF

Expresadas las estimaciones de los parámetros, obtenidas a partir de los grupos de comparación, en una métrica común y elegido un procedimiento de purificación de la escala, la estrategia basada en la TRI para identificar los ítems posiblemente sesgados consiste en la comparación de las estimaciones para los grupos de interés. El punto clave es entonces la identificación del procedimiento que conduzca a una comparación más fina y precisa. Las

[5] Uno de los métodos más utilizado para la detección de DIF por su adecuada eficiencia y bajo costo es el MantelHaenszel; algunos de los trabajos más recientes sobre este procedimiento son Fidalgo, Ferreres & Muñiz (2004a); Fidalgo, Ferreres & Muñiz (2004b); Hidalgo Montesinos & López Pina (2004) y Penfield (2001). Para revisiones generales sobre diferentes métodos pueden verse Gómez Benito & Hidalgo Montesinos (1997) y Herrera, Sánchez Pedraza & Gómez Benito (2001)

diferentes propuestas metodológicas se han agrupado en tres: comparación de los modelos ajustados, comparación de los parámetros de los mismos y medidas del área de discrepancia entre las CCI.

Comparación de modelos

La primera referencia que generalmente se cita sobre la propuesta de comparación de modelos TRI para la detección de DIF es la aplicación de Thissen, Steinberg & Gerrard (1986). Los autores ilustraron un procedimiento para probar la hipótesis de ausencia de DIF para uno o varios ítems simultáneamente, evaluando el ajuste de modelos TRI para los grupos de interés, mediante el estadístico G^2 de Bishop, Fienberg & Holland (1975). Posteriormente Thissen et al., (1988) y Thissen et al., (1993) formalizaron e ilustraron más ampliamente el procedimiento. Brevemente, éste consiste en la comparación del ajuste de dos modelos, uno de los cuales, generalmente denominado modelo compacto (C), supone que los parámetros son iguales para los dos grupos, y otro, el modelo aumentado (A), supone que tales parámetros son diferentes. La significación de las diferencias observadas entre los modelos se evalúa mediante el estadístico de razón de verosimilitud. Thissen et al., (1988, p. 153) describen el procedimiento en tres pasos: a) ajustar un modelo TRI (modelo A) para los dos grupos simultáneamente con uno o unos pocos ítems 'de anclaje', a los cuales se le impone la restricción de que los parámetros sean iguales para los dos grupos; esta restricción no se impone al ítem o ítems estudiados; una vez estimados los parámetros se calcula el estadístico $G_A^2 = -2 \log L(A)$, donde *L(A)* es la función de verosimilitud para las estimaciones de los parámetros del modelo, b) ajustar nuevamente el modelo (modelo C) imponiendo la restricción de igualdad de parámetros para el o los ítems estudiados y calcular $G_C^2 = -2 \log L(C)$, y c) calcular el estadístico $G^2 = G_C^2 - G_A^2$ para probar la hipótesis de igualdad de los modelos o ausencia de DIF; éste sigue una distribución χ^2 con tantos grados de libertad como la diferencia en el número de parámetros de los dos modelos. El rechazo de la hipótesis de igualdad de los modelos, conduce a la conclusión de que el ítem o ítems analizados presentan DIF.

Otras propuestas de procedimientos que siguen el mismo razonamiento son la de Muthén & Lehman (1985), la de Kelderman (1989) y la de Bock, Muraki & Pfeiffenberger (1988). La primera de ellas estima los parámetros mediante mínimos cuadrados generalizados para modelos TRI basados en la distribución normal acumulada, y posteriormente evalúa la significación de las diferencias entre los modelos mediante la razón de verosimilitud ya mencionada. Kelderman (1989) por su parte propuso un procedimiento para comparar modelos loglineales de un parámetro estimando los parámetros por máxima verosimilitud y sometiendo a prueba la hipótesis de no DIF mediante el mismo estadístico G^2. Finalmente, Bock et al.,(1988) propusieron estimar los parámetros por máxima verosimilitud marginal y evaluar las diferencias observadas utilizando el error estándar de dichas estimaciones. Fidalgo (1996) incluye en su texto una ilustración detallada del procedimiento de Thissen et al.,(1986) y en Thissen et al., (1993) se encuentran ejemplos numéricos de cada uno de los cuatro procedimientos.

Aunque estos procedimientos gozan de un adecuado sustento matemático y permiten evaluar una prueba completa o un grupo de ítems de manera simultánea, no se han generalizado en la

práctica. Según Gómez Benito & Hidalgo Montesinos (1997) las dos grandes desventajas del procedimiento de Thissen et al., (1986) son su incapacidad para detectar diferencias pequeñas y su dependencia del tamaño de los grupos, mientras que el de Kelderman (1989) presenta dificultades para análisis simultáneos de pruebas largas y requiere demasiadas iteraciones para alcanzar un criterio de convergencia. Presumiblemente una razón válida para el escaso uso actual de todos los procedimientos descritos es su alto costo computacional en comparación con otros procedimientos disponibles hoy, incluyendo los demás basados en la TRI. De acuerdo con Kim & Cohen (1995) utilizando el procedimiento de Thissen et al., (1986) y Thissen et al.,(1988, 1993) iterativamente, se requieren $3k+2ni-i^2$ calibraciones para el análisis de DIF de los ítems, donde *k* es el número de iteraciones necesarias y *i* es el número de ítems; así, con 3 iteraciones y 15 ítems (prueba más corta de lo habitual), el número de calibraciones sería 90. Además, aunque no se utilice el procedimiento iterativo y se recurra a otro mecanismo de purificación o si no se utiliza purificación alguna, puesto que requiere múltiples calibraciones, el procedimiento resulta más dispendioso que otros basados en la TRI y definitivamente mucho más que otros como los basados en tablas de contingencia.

Comparación de parámetros: χ^2 de Lord

La propuesta pionera en las técnicas de detección de DIF con base en modelos TRI es la de Lord (1980) presentada en su libro sobre aplicaciones de los modelos TRI, la cual toma parte de un trabajo anterior (Lord, 1977). Este procedimiento compara los vectores de los parámetros estimados cuando se ajustan modelos TRI para los dos grupos separadamente. Brevemente el procedimiento consiste en estimar los parámetros de los ítems ajustando modelos TRI para los grupos de interés, poner las estimaciones en la misma métrica y someter a prueba estadística las diferencias observadas entre los vectores de estimaciones de los parámetros.

Si X_1 y X_2 son los vectores de parámetros para los dos grupos, la hipótesis de ausencia de DIF, puede expresarse como $H_0 : X_1 = X_2$. Ahora, si $\hat{X}_1$ y $\hat{X}_2$ son los vectores de las estimaciones de los mismos parámetros, las diferencias entre dichas estimaciones para los dos grupos pueden expresarse en un vector $V = \hat{X}_1 - \hat{X}_2$ que tiene dimensión *p X 1*, donde *p* es el número de parámetros en el modelo; así, si se ajusta modelos de 1P, $V = \hat{b}_{i1} - \hat{b}_{i2}$. El estadístico para someter a prueba la significación de las diferencias así observadas, conocido como el χ^2 de Lord, es $\chi^2 = V' \sum^{-1} V$, donde $\sum$ es la matriz de varianzas y covarianzas de las diferencias entre los parámetros. Esta matriz se estima como la suma de las matrices de varianzas y covarianzas de las estimaciones de los parámetros para los dos grupos; esto es: $S = S_1 + S_2$, donde S_1 y S_2 son las matrices de varianzas y covarianzas de las estimaciones de los parámetros para cada grupo. En estos términos el estadístico de prueba sigue una distribución χ^2 con *p* grados de libertad, tantos como parámetros comparados, y queda expresado como $\chi^2 = V' S^{-1} V$. Lord (1980) recomienda que las estimaciones de los parámetros se realicen estandarizando las del parámetro de dificultad en una escala con media *0* y desviación estándar *1*, con el fin de que las estimaciones de todos los parámetros en todos los grupos queden expresadas en la misma escala. Además, para el cálculo de la media y desviación estándar que se utilizan para dicha estandarización, puede ser conveniente omitir los ítems con dificultad extrema (muy fáciles o muy difíciles) ya que las estimaciones de este parámetro para este tipo de ítems suelen tener error

grande; en todo caso, estos ítems se omiten únicamente para el cálculo de dichos descriptivos y reciben el mismo tratamiento que los otros ítems, para todos los demás efectos.

Lord (1980, p. 217) advirtió que cuando se ajustan modelos de tres parámetros y los ítems difieren en el parámetro de discriminación, los más discriminativos mostrarán mayores diferencias entre los grupos comparados; considerando además que el DIF se expresa a través de las diferencias entre la dificultad y la discriminación, el uso del parámetro de pseudoazar no parece relevante y en cambio presenta dificultades en su estimación. De esta manera el autor sugiere un procedimiento de tres pasos cuando se utilicen modelos de tres parámetros, así: a) ajustar un único modelo de tres parámetros tomando conjuntamente los grupos que se van a comparar, para obtener una estimación conjunta del parámetro c_i, b) ajustar modelos de tres parámetros separadamente para los dos grupos, fijando como valor del parámetro c_i, la estimación obtenida en el paso anterior, y c) someter a prueba mediante el estadístico χ^2, la hipótesis de diferencia de los parámetros de dificultad y discriminación, con base en las estimaciones obtenidas en el paso anterior.

Además de la limitación del procedimiento cuando se ajustan modelos de tres parámetros, los supuestos que lo sustentan han impuesto otras limitaciones que pueden ser serias en sus aplicaciones prácticas. En primer lugar, la prueba χ^2 de Lord es asintótica y en consecuencia hay una exigencia de tamaño de muestra para alcanzar la distribución esperada, exigencia que puede ser difícil de satisfacer cuando se trabaja con grupos minoritarios; de otra parte, supone que el parámetro de magnitud de atributo de los individuos (θ) es conocido, supuesto imposible de cumplir en los estudios aplicados; y finalmente, es aplicable únicamente cuando se utilizan algoritmos de estimación por máxima verosimilitud. Estas limitaciones, el desconocimiento de una tamaño de muestra mínimo para lograr la convergencia a la distribución χ^2 y algunos reportes como el de Linn, Levine, Hastings & Wardrop (1981) según el cual podía presentar una alta tasa de falsos positivos (FP) en comparación con medidas de área, hicieron que el procedimiento adquiriera mala prensa. De hecho, Camilli & Shepard (1994), en uno de los textos más consultados sobre métodos para la detección de DIF, no recomiendan su uso; otros autores como Hidalgo Montesinos & López Pina (1997) y López Pina, Hidalgo & Sánchez Meca (1993) no lo prefieren comparado con procedimientos no basados en la TRI y Fidalgo, (1996) y Millsap & Everson (1993) sugieren utilizarlo en combinación con otros procedimientos, sobre todo si resulta significativo.

Sin embargo, buena parte de los trabajos posteriores a la publicación de la propuesta de Lord (1980) se han dedicado al examen del funcionamiento del estadístico en condiciones diferentes a las supuestas por él. Con respecto al algoritmo de estimación de los parámetros y el supuesto θ conocido, los estudios han sido bastante consistentes en mostrar que las estimaciones por máxima verosimilitud marginal (MVM) y bayesiana (MB) pueden mejorar sustancialmente la precisión de los resultados y por tanto, el funcionamiento del χ^2 de Lord. McLaughlin & Drasgow (1987) mostraron que cuando los parámetros de los ítems y de los individuos son todos desconocidos y se estiman mediante máxima verosimilitud conjunta (MVC), el χ^2 de Lord presenta una importante inflación del error tipo I, por encima de los valores nominales; Kim & Cohen (1994)

replicaron este estudio utilizando estimaciones MVM y MB con modelos de 2 y 3 parámetros y encontraron que el error tipo I se mantuvo satisfactoriamente controlado con modelos de 2 parámetros y con 3 parámetros si se fija un valor para *c* (3P-c). Los mismos autores (Cohen & Kim, 1993) habían encontrado que las tasas de falsos positivos y falsos negativos fueron más satisfactorias cuando se utilizó el algoritmo de estimación bayesiana. Por su parte, Lim & Drasgow (1990) encontraron que las estimaciones MVM y MB arrojaron resultados satisfactorios con ítems unidimensionales y con tamaños de grupos de 750 examinados, además, encontraron cierta superioridad de las estimaciones bayesianas sobre las de MVM con grupos pequeños (250 examinados por grupo).

En cuanto al efecto del tamaño de muestra sobre el funcionamiento del χ^2 de Lord para la detección de DIF, en el mismo estudio anterior, Lim & Drasgow (1990) advirtieron que tanto las estimaciones por MVM como la MB son sensitivas al tamaño de muestra y ese efecto es mayor para las estimaciones por máxima verosimilitud. Además, con modelos de 2 parámetros, Cohen & Kim (1993) recomendaron su uso cuando se trabaja con muestras pequeñas (100 examinados por grupo), pruebas cortas (hasta 20 ítems), diferentes distribuciones de la magnitud de atributo o altos porcentajes de ítems con DIF (hasta 20%). Finalmente, otro aspecto que se ha mencionado como una debilidad del procedimiento es la estimación de la matriz de varianzas y covarianzas, necesaria para el cálculo del estadístico de prueba. Según Thissen & Wainer (1982) la falta de precisión en la estimación de la matriz de varianzas y covarianzas puede aumentar el error en la detección de ítems con DIF; sin embargo, Kim & Cohen (1995) concluyeron que "la carencia de las covarianzas fuera de la diagonal de matriz de varianzas y covarianzas, no parece afectar de manera importante el acuerdo entre esas medidas en la detección de DIF" (p. 309), además, hallaron tasas de error dentro de los límites nominales.

Finalmente, debe anotarse que comparado con otros procedimientos basados en la TRI, el χ^2 es un procedimiento relativamente fácil de utilizar, cuenta con una prueba de significación para la identificación de los ítems con DIF, y algunos estudios apoyan su uso en la práctica. Con datos simulados utilizando modelos TRI de dos parámetros McCauley & Mendoza (1985) encontraron que el χ^2 de Lord resultó más efectivo que otros índices basados en la TRI, incluyendo los de medidas de área, para identificar ítems con DIF cuando éste se ha simulado generando los datos con base en modelos multidimensionales. Los resultados del estudio de Kim & Cohen (1995) con respuestas a una prueba de matemáticas y ajustando modelos de dos parámetros también apoyan su uso sobretodo cuando se compara en términos de costo computacional con la razón de verosimilitud de Thissen et al.,(1988, 1993). Evaluando la efectividad del proceso de purificación de dos etapas sobre la efectividad de métodos basados en la TRI, Núñez Núñez, Hidalgo Montesinos & López Pina (2000) encontraron resultados satisfactorios para el χ^2 de Lord, en comparación con medidas de área. Hidalgo Montesinos & López Pina (2002) aplicaron un procedimiento de purificación de dos etapas con modelos de respuesta graduada utilizando una medida de área y el χ^2 de Lord y no encontraron una superioridad notoria de alguno de los dos procedimientos, aunque hacen notar que el último es un estadístico más conservador que la medida de área (p. 43). Nótese, sin embargo, que en estos estudios se obtuvieron estimaciones de máxima verosimilitud marginal y Cohen & Kim (1995) utilizaron también estimaciones bayesianas; además, en ningún caso se ajustaron modelos TRI de tres parámetros.

Medidas de área entre las CCI

Aunque las medidas de área parten de la misma idea básica común a todos los procedimientos basados en la TRI, evalúan el área comprendida entre las dos CCI del ítem ajustadas para los dos grupos independientemente y expresadas en la misma métrica, así, un ítem estará libre de DIF si dicha área es nula. La primera propuesta en este sentido fue la de Rudner (1977) y Rudner, Getson & Knight (1980a, 1980b), que se conoce hoy como una aproximación discreta para un intervalo finito de θ, y aparece ilustrada en la figura 2. Después de ajustar los modelos y equiparar las dos CCI con valores de θ en un intervalo fijo (generalmente $-3 \le \theta \le 3$), esta escala se divide en pequeños intervalos ($\Delta\theta$) y se calcula el área del rectángulo que mide por un lado $\Delta\theta$, y el valor absoluto de la diferencia de probabilidades entre los grupos dado un valor θ_j ($\left|P_1(\theta=\theta_j)-P_2(\theta=\theta_j)\right|$), por el otro lado. El índice de discrepancia es entonces la suma de las áreas para todo el rango de valores de θ ($A=\sum_{\forall\theta_j}\left|P_1(\theta=\theta_j)-P_2(\theta=\theta_j)\right|\Delta\theta$). En la figura 2 se ilustra el cálculo de este índice para un ítem con parámetros $a_r = 1.7$, $b_r = -.65$, $a_f = 1$, $b_r = .2$ y $c_r = c_f = 0$ haciendo $\Delta\theta = .1$, valor útil para efectos ilustrativos pero demasiado grande para un estudio aplicado puesto que la precisión del índice depende de la amplitud del incremento.

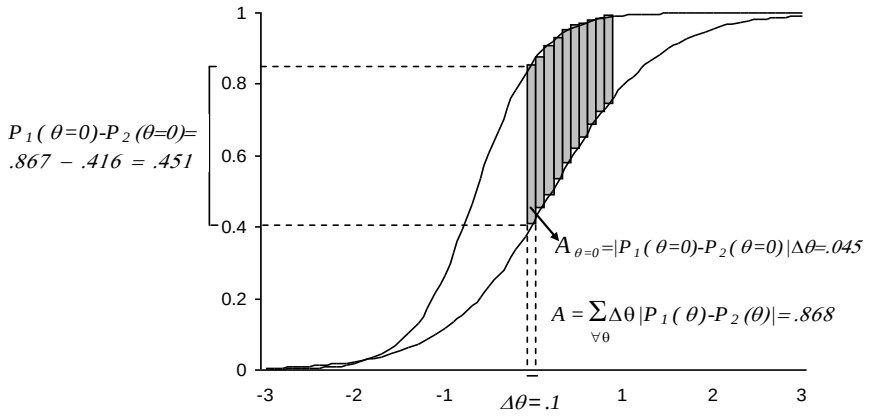

Figura 2: Ilustración de la medida de área sin signo propuesta por Rudner, (1977) y Rudner et al., (1980a, 1980b) con $\Delta\theta=.1$

Algunas de las críticas a este índice sencillo, muy intuitivo y fácil de calcular, fueron la ausencia de un estadístico de prueba o un punto de corte que permitiera clasificar los ítems en sesgados e insesgados, su posible imprecisión y dar el mismo peso a las áreas en todos los puntos de la escala de θ. Varias propuestas posteriores han intentado superar estas limitaciones, algunas de ellas son los cuatro índices de Linn et al., (1981), los índices de diferencia de probabilidad de Linn & Harnisch (1981), la suma de cuadrados autoponderada de Shepard, Camilli & Williams (1984) o la medida de áreas con signo de Camilli & Shepard (1994). Aunque algunas de éstas logran ponderar las diferencias teniendo en cuenta el número de examinados en cada intervalo y en consecuencia, mejorarían la precisión del cálculo, Shepard, Camilli & Williams (1985) no encontraron diferencias importantes entre las medidas ponderadas y las no ponderadas; además,

todas las propuestas siguen adoleciendo de un valor crítico para identificar la discrepancia máxima atribuible al azar y son medidas aproximadas puesto que calculan áreas discretas para evaluar un área que es de naturaleza continua.

Desde una perspectiva matemática diferente, Raju (1988) propuso unas medidas exactas que toman en consideración la naturaleza continua de la escala de θ, integrando entre $-\infty$ e ∞ con respecto a θ. Si $F_1(\theta)$ y $F_2(\theta)$ son las funciones de respuesta al ítem para los dos grupos, el área exacta signada (ESA, por las siglas en inglés) y no signada (EUA) entre las respectivas CCI son $ESA = \int_{-\infty}^{+\infty} (F_1(\theta) - F_2(\theta)) d(\theta)$ y $EUA = \int_{-\infty}^{+\infty} |F_1(\theta) - F_2(\theta)| d(\theta)$, respectivamente. Con estos puntos de partida, Raju (1988), derivó las fórmulas para calcular las medidas de área signadas y no signadas para modelos de 1, 2 y 3 parámetros; sin embargo cuando se ajustan modelos 3P, el cálculo del área es posible si *c* es igual para los dos grupos, si esta condición no se cumple, el área es infinita (p. 499). Con esta restricción, las medidas de área signada y no signada entre CCI para modelos de tres parámetros, son $ESA = (1-c)(b_1 - b_2)$ y $EUA = (1-c)\left| \frac{2(a_1 - a_2)}{da_1 a_2} ln\left(1 + e^{\frac{da_1 a_2 (b_1 - b_2)}{a_1 - a_2}}\right) - (b_1 - b_2) \right|$, respectivamente. Si además el ítem es igualmente discriminativo para los dos grupos ($a_1 = a_2 = a$), ésta última expresión se reduce a $EUA = (1-c)|b_1 - b_2|$. Las expresiones para modelos de dos parámetros ignoran los términos que dependen de *c*, y para modelos de 1P se reducen a la diferencia entre la dificultad: $ESA = b_1 - b_2$ y $EUA = |b_1 - b_2|$.

Aunque el mismo autor hizo notar que las diferencias entre las CCI son más importantes en los valores de θ con mayor número de examinados y en consecuencia puede ser apropiado calcularlas para intervalos cerrados, mostró que la diferencia entre las dos aproximaciones puede ser numéricamente importante y que algunos ítems detectados con las medidas de área exacta pueden parecer no sesgados con intervalo cerrado (ver ejemplo en Raju, 1980, p. 501). Dos años más tarde el mismo autor (Raju, 1990) describió las distribuciones muestrales de sus medidas de área y propuso los estadísticos de prueba para *ESA* y *EUA*, conocidos como *Z(ESA)* y *Z(H)*, respectivamente, asintóticamente normales. La primera está definida para modelos de 1P y para modelos 2P y 3P siempre y cuando tengan la misma discriminación para los dos grupos[6], y tiene la forma general $Z(ESA) = \frac{\hat{b}_1 - \hat{b}_2}{\sqrt{Var(ESA)}}$, donde la varianza se calcula dependiendo del tipo área (signada o no signada) y del modelo específico. La segunda prueba está definida para modelos de dos y tres parámetros con diferente discriminación y tiene la forma $Z(H) = \frac{H}{\sqrt{Var(H)}}$, donde *H* y su varianza se calculan

[6] Si los parámetros de discriminación son iguales para los dos grupos ($a_1 = a_2$) las CCI no se cruzan, mientras que si $a_1 \neq a_2$ éstas se cruzan en un punto de la escala de θ.

dependiendo del tipo de modelo. Las expresiones para calcular tales valores pueden consultarse en Raju (1990) o Fidalgo (1996).

Partiendo del mismo punto pero integrando para un intervalo cerrado entre θ_1 y θ_2, Kim & Cohen (1991) derivaron las expresiones para calcular las medidas de área entre las CCI, que se conocen como medidas de área cerradas signadas (CSA, por sus siglas en inglés) y no signadas (CUA) para modelos de 1P, 2P y 3P. En este caso no existe la limitación de las áreas de Raju en los modelos de 3P; puesto que se limita el intervalo de θ, el área entre las CCI es finita aún cuando los modelos sean de tres parámetros con diferente valor de *c*. Para modelos de 1P, modelos de 2P con igual discriminación y modelos de 3P con igual discriminación y pseudoazar, el área no signada es sencillamente el valor absoluto del área signada respectiva, es decir $CUA = |CSA|$. Sin embargo, cuando estas condiciones no se cumplen y las dos CCI se cruzan en un punto θ_x, el cálculo del área no signada depende de la localización de dicho punto: Si este valor se encuentra fuera del intervalo ($\theta_x < \theta_1$ o $\theta_x > \theta_2$) el área no signada sigue siendo el valor absoluto de la signada, pero si las CCI se cruzan en un punto que cae dentro del intervalo de interés ($\theta_1 < \theta_x < \theta_2$), el cálculo del área debe considerar dicho valor. Las expresiones para obtener el valor θ_x y para el cálculo de todas estas medidas de área se encuentran en Kim & Cohen (1991). En el mismo estudio los autores encontraron resultados muy similares entre las medidas de área exactas y cerradas con modelos de 3P y 3P-c; además sus hallazgos mostraron que éstas últimas funcionaron satisfactoriamente con ambos tipos de modelos, este resultado apoyaría su uso en comparación con las medidas exactas puesto que no hacen la restricción de igualdad del parámetro *c*.

A pesar de su sencillez, una revisión de los estudios publicados indican que actualmente las medidas de área parecen ser más utilizadas para simular el DIF que para detectarlo; además, algunos resultados de estudios comparativos no favorecen su uso. De acuerdo con los resultados de Cohen y Kim (1993), aunque las diferencias no son demasiado grandes, tanto el *Z(ESA)* como el *Z(H)* mostraron resultados más pobres que el χ^2 de Lord. Ajustando modelos de 1P, Gómez Benito & Navas-Ara (2000) encontraron resultados menos satisfactorios para las medidas de área exactas que para otros procedimientos no basados en la TRI y Núñez Núñez, Hidalgo Montesinos & López Pina (2000) encontraron resultados menos satisfactorios para las medidas de área en comparación con el χ^2 de Lord.

Conclusiones

El presente trabajo ha presentado una revisión de las principales publicaciones de las últimas décadas que se refieren al diseño o evaluación de propuestas metodológicas para la detección de DIF, generadas en el marco de la TRI. Esta revisión se ha limitado a la detección de ítems dicótomos mediante el ajuste de modelos unidimensionales cuando se comparan solamente dos grupos, condiciones muy frecuentes en la práctica profesional; sin embargo, debe anotarse que también existen importantes desarrollos metodológicos para la identificación de ítems de respuesta graduada (Cohen et al., 1993; Kim & Cohen, 1998, e Hidalgo Montesinos & López Pina, 2002, entre muchos otros) y cuando interesa comparar más de dos grupos (Kim, Cohen & Park, 1995; Muthén & Lehman, 1985 y Penfield, 2001). Además, dentro de los trabajos sobre el

desarrollo de modelos TRI multidimensionales pueden citarse los de Reckase (1997) y McDonald (1997, 2000); mientras que Shealy & Stout (1993b) y Bolt (1996) presentan aplicaciones de este tipo de modelos en la detección de DIF. Tanto las especificidades de los modelos TRI multidimensionales o con ítems no dicótomos, como sus aplicaciones en la detección de DIF merecen una exhaustiva revisión que trasciende los objetivos de la presente.

Algunas de las propuestas metodológicas desarrolladas dentro de la TRI para la detección de DIF se ocupan de los procedimientos necesarios para garantizar que los grupos comparados sean realmente comparables mientras que otros centran su atención en los procedimientos que prueban las hipótesis de ausencia de DIF. Tanto los procedimientos de equiparación de puntuaciones como de purificación de la escala de θ, constituyen temas de candente actualidad; sin embargo, los trabajos que evalúan la eficiencia relativa de los procedimientos de estimación de las constantes de equiparación no permiten extraer resultados concluyentes, algunos apoyan el uso del método de función de respuesta al ítem o el χ^2 mínimo, otros lo encuentran menos recomendable que los procedimientos no basados en la TRI y otros no encuentran diferencias entre las diversas propuestas. Además, una somera revisión de publicaciones actuales sobre trabajos aplicados en la detección de DIF, tampoco parece favorecer a unos u otros puesto que no se observa una clara predominancia en su uso.

De otra parte, una de las conclusiones más evidentes y nada novedosa, es que resulta recomendable el uso de algún procedimiento de purificación de la escala para mejorar la precisión en la detección de ítems con DIF; sin embargo, uno de los aspectos que pesa a la hora de elegir un procedimiento es el costo computacional que se incrementa de manera importante con el número de calibraciones necesarias. En ese sentido el procedimiento bietápico de Hidalgo Montesinos & López Pina (2002) puede representar una opción llamativa por su economía, si la investigación posterior muestra su eficiencia. Finalmente, en lo que tiene que ver con esta primera categoría de trabajos llama la atención los recientes cuestionamientos a la práctica de conformación de la prueba de anclaje en los estudios de DIF. Dado que en este tipo de estudios se tiene la misma prueba aplicada a grupos diferentes, resulta razonable utilizar como subprueba de anclaje el grupo de ítems de la prueba exceptuando el que se está estudiando; sin embargo, las investigaciones muestran que la utilidad de dichas subpruebas en términos de la precisión de la estimación de las constantes de equiparación, no depende solamente de su longitud sino del grado en que se encuentre libre de sesgo. Trabajos como los de Wang & Yeh (2003) y Wang, (2004) sugieren que unos pocos ítems libre de DIF pueden conformar una subprueba de anclaje muy adecuada; sin embargo, queda aún abierta la discusión en torno al número mínimo necesario y las propuestas procedimentales para su elección en los estudios de DIF.

Con respecto a los métodos de detección de DIF, aunque la revisión no permite conclusiones definitivas, el χ^2 parece ser comparativamente preferible. A los procedimientos de comparación de modelos se le reconoce su adecuado y elegante sustento matemático y su capacidad para evaluar simultáneamente una prueba completa o un grupo de ítems; sin embargo, se les acusa de muy exigente desde el punto de vista de costo computacional sobre todo cuando se utilizan procedimientos de purificación de la escala (Gómez Benito & Hidalgo Montesinos, 1997; Kim & Cohen (1995). De otra parte, aunque las medidas de área entre las CCI parecen bastante intuitivas y más económicas que las anteriores, los resultados tampoco son consistentes en favorecer su uso en comparación con el χ^2 de Lord (Núñez Núñez, Hidalgo Montesinos & López Pina, 2000) o con

otros procedimientos no basados en la TRI (Gómez Benito & Navas-Ara, 2000). La mayor limitación en las aplicaciones reales es su exigencia en el tamaño de muestra necesario para obtener estimaciones adecuadas de los parámetros de los modelos y el cálculo de las constantes de equiparación. Aunque algunos trabajos como los de Lim & Drasgow (1990) o Cohen & Kim (1993) reportan resultados satisfactorios de las estimaciones bayesianas con grupos pequeños, la práctica más generalizada en la detección de DIF consiste en elegir procedimientos no basados en la TRI como el Mantel-Haenszel, comparación de proporciones o el delta-plot (Muñiz, Hambleton & Xing, 2001).

Finalmente, vale la pena mencionar que de acuerdo con Kim & Cohen (1995) existe un buen acuerdo entre los tres procedimientos basados en la TRI (razón de verosimilitud, χ^2 de Lord y medidas de área *Z(ESA)* y *Z(H)*) en la detección de DIF. Este acuerdo y las similitudes entre los resultados arrojados por los tres procedimientos mejora cuando se utilizan procedimientos iterativos; a partir de estos resultados los autores recomiendan usar algún procedimiento de purificación de la escala y no utilizar una única técnica en las aplicaciones prácticas, sino usar una combinación de ellas. De acuerdo con Hambleton, Clauser, Mazor & Jones (1993), y Gómez Benito & Hidalgo Montesinos (1997) los métodos basados en la TRI que estiman un nivel de habilidad o atributo latente gozan hoy de mucha aceptación, entre sus bondades se destacan generalmente su solidez matemática y su capacidad para detectar DIF uniforme y no uniforme.

Referencias

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

Angoff, W. H. (1972, september). *A technique for the investigation of cultural differences*. Documento presentado en el Annual Meeting of the American Psychological Association: Honolulu.

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, N. J.: Educational Testing Service.

Angoff, W. H. & Ford, S. F. (1973). Item race interaction on a test of scholastic aptitude. *Journal of Educational Measurement. 10*, 95-105.

Baker, F. B. (1995). *Equate computer Program Version 2.1*. Madison, Wisconsin: Laboratory of Experimental Design. Department of Educational Psychology. University of Wisconsin.

Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. *Applied Psychological Measurement. 20*, 45-57.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement. 28*, 147-162.

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

Binet, A. & Simon, T. (1916). *The development of intelligence in children*. New York: Amo.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a examinee's ability. En F.M.Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (chapter 2). Reading, Mass: Addison-Wesley.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge: MIT Press.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika. 25*, 179-197.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika. 46*, 443-449.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement. 25*, 275-285.

Bolt, D. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika. 23*(1), 67-95.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (Ed) (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement. 26* (4), 433-450.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. 4. Newbury Park, C.A: SAGE Publications.

Candell, G. & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement. 12*(3), 253-260.

Cohen, A. S. & Kim, S.-H. (1993). A Comparison of Lord´s X^2 and Raju´s Area Measures in detection of DIF. *Applied Psychological Measurement. 17*(1), 39-52.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement. 17*(4), 335-350.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando: Harcourt Brace Jovanovich College Publishers.

Cuesta, M. (1996). Unidimensionalidad. En J.Muñiz (Ed.), *Psicometría* (pp. 239-291). Madrid: Editorial Universitas S.A.

Cuesta, M. & Muñiz, J. (1994). Utilización de los modelos de teoría de respuesta a los ítems con datos multifactoriales. *Psicothema. 6*(2), 283-296.

Cuesta, M. & Muñiz, J. (1995). Efectos de la multidimensionalidad en la estimación de parámetros desde modelos unidimensionales de teoría de respuesta a los ítems. *Psicológica. 16*(1), 65-86.

DeMars, C. (2000). DRAWICC: Modules to Graph Item Response Functions and Item Information Functions with SAS GPLOT. *Applied Psychological Measurement. 24*(3), 224.

Divgi, D. R. (1985). A minimum chi-square methods for developing a common metric in IRT. *Applied Psychological Measurement. 9*(4), 413-415.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education. 2*, 217-233.

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

Drasgow, F. (1987). Study of the Measurement Bias of Two Standardized Psychological Tests. *Journal of Applied Psychology. 72*(1), 19-29.

Drasgow, F. & Parsons, C. K. (1993). Aplication of unidimensional item response theory model to multidimensional data. *Applied Psychological Measurement. 7*(2), 189-199.

Eelles, K., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and cultural differences*. Chicago: University of Chicago Press.

Fidalgo, Á. M. (1996). Funcionamiento Diferencial de los Items. En J.Muñiz (Ed.), *Psicometría* (pp. 371-455). Madrid: Editorial Universitas, S.A.

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004a). Liberal and Conservative Differential Item Functioning Detection Using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II Error Rates. *The Journal of Experimental Education. 7*(I), 23-39.

Fidalgo, Á. M., Ferreres, D., & Muñiz, J. (2004b). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement. 64*(6), 925-936.

Fox, J. P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology. 56*, 65-81.

Gómez Benito, J. & Hidalgo Montesinos, M. D. (1997). Evaluación del funcionamiento diferencial en items dicotómicos: una revisión metodológica. *Anuario de Psicología.(74)*, 3-32.

Gómez Benito, J. & Navas-Ara, M. J. (2000). A Comparison of $X^{2,\ \text{RFA and IRT Based Procedures in the Detection of DIF}}$. *Quality & Quantity. 34*, 17-31.

Gómez-Benito, J. & Hidalgo-Montesinos, M. D. (2003). Desarrollos recientes en psicometría. *Avances en Medición. 1*, 17-36.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review. 9*, 139-150.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications.

Hanson, B. A. (2001). Estimation Program for Dichotomous Item Response Models (EPDIRM) [Computer Software].

Harvey, R. & Hammer, A. (1999). Item Response Theory. *Counseling Psychologist. 27*(3), 353-383.

Herrera, A. N., Sánchez Pedraza, R., & Gómez Benito, J. (2001). Funcionamiento diferencial de los Items: Una revisión conceptual y metodológica. *Acta Colombiana de Psicología. 5*, 41-61.

Hidalgo Montesinos, M. D. & López Pina, J. A. (1997). Comparación entre las medidas de área, el estadístico de Lord y el análisis de regresión logística en la evaluación del funcionamiento diferencial de los ítems. *Psicothema. 9*, 417-431.

Hidalgo Montesinos, M. D. & López Pina, J. A. (2002). Two-stage equating differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement. 62*, 32-44.

Hidalgo Montesinos, M. D. & López Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement. 64*(6), 903-915.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. En H.Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, N.J.: Erlbaum.

Holland, P. W. & Wainer, H. (1993). Preface. En P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning*, New Jersey: Lawrence Erlbaum Associates, Inc.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow-Jones-Irwin.

Jensen, A. R. (1969). How much can we boast IQ and scholastic achievement? *Harvard Educational Review. 39*, 1-123.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Kaskowitz, G. S. & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function methods of linking. *Applied Psychological Measurement. 25*(1), 39-52.

Kelderman, H. (1989). Item Bias Detection Using Loglinear IRT. *Psychometrika. 54*(4), 681-697.
Kim, S.-H. & 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. & Cohen, A. S. (1992). Effects on linking methods on detection of DIF. *Journal of Educational Measurement. 29*, 51-66.
Kim, S.-H. & Cohen, A. S. (1994). An investigation of Lord´s Procedure for the detection of Diferential Item Functioning. *Applied Psychological Measurement. 18*(3), 217-228.
Kim, S.-H. & Cohen, A. S. (1995). A comparison of Lord´s Chi-Square, Raju´s Area Measures, and the Likelihood Ratio Test on detection of Differential Item Functioning. *Applied Measurement in Education. 8*(4), 291-312.
Kim, S.-H. & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio Test. *Applied Psychological Measurement. 22*(4), 345-355.
Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement. 32*(3), 261-276.
Kok, F. (1988). Item bias and test multidimensionality. En R.Langeheine & J. Rost (Eds.), *Latent trait and latent class models*, (pp. 263-274). New York: Plenum.
Kolen, M. J. (2004). Linking Assessments: Concept and History. *Applied Psychological Measurement. 28*(4), 219-226.
Lautenschlager, G., Flaherty, V. L., & Park, D.-G. (1994). IRT Differential Item Functioning: An Examination of Ability Scale Purifications. *Educational and Psychological Measurement. 54*(1), 21-31.
Lautenschlager, G. & Park, D.-G. (1988). IRT Item bias detection procedures; Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement. 12*(4), 365-376.
Lawley, D. N. (1943). On problems connected with ítem selection and test construction. *Proceedings of the Royal Society of Edimburg. 61*, 273-287.
Lazarsfeld, P. F. (1950). The logical and mathamatical foundation of latent structure analysis. En S.A.Stoufer (Ed.), *Measurement and Prediction*, Princeton: Princeton University Press.
Lim, R. G. & Drasgow, F. (1990). Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning. *Journal of Applied Psychology. 75*(2), 164-174.
Linacre, M. (2003). *A user's guide to Winsteps*. Chicago: Mesa Press.
Linn, R. L. & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement. 18*, 109-118.
Linn, R. L., Levine, M. V., Hastings, G. N., & Wardrop, J. L. (1981). Item Bias in a test of reading comprehension. *Applied Psychological Measurement. 5*, 159-173.
López Pina, J. A., Hidalgo, M. D., & Sánchez Meca, J. (1993). Error tipo I de las pruebas Chi-cuadrado en el estudio del sesgo de los items. En C.Arce & G. Seoane (Eds.), *III Simposium de metodología de las ciencias sociales y del comportamiento* (pp. 521-529).
Lord, F. M. (Ed) (1952). A theory of test scores. *Psychometric Monographs.* (7).
Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. En Y.H.Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets and Zeitlinger.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J.: Lawrence Erlbaum Associates.
Lord, F. M. & Novick, M. R. (1966). *Statistical theories of mental test scores.* Reading, MA: Adison Wesley.
Marco, G. L. (1977). Item Characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement. 14*(2), 139-160.
McCauley, C. D. & Mendoza, J. (1985). A Simulation Study of Item Bias Using a Two-parameters Item Response Model. *Applied Psychological Measurement. 9*(4), 389-400.
McDonald, R. P. (1997). Normal-Ojiva multidimensional model. En W.J.van der Linder & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 258-269). New York: Springer.
McDonald, R. P. (2000). A basic for multidimensional item response theory. *Applied Psychological Measurement. 24*(2), 99-114.
McLaughlin, M. E. & Drasgow, F. (1987). Lord´s chi-square test of item bias with estimated and with know person parameters. *Applied Psychological Measurement. 11*, 161-173.
Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics. 7*, 105-118.
Miller, M. D. & 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*(4), 381-388.
Millsap, R. & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement. 17*(4), 297-334.
Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika. 51*, 177-195.
Mislevy, R. J. & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville IN: Scientific Software.
Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Items.* Madrid: Ediciones Pirámide.
Muñiz, J. & Hambleton, R. K. (1992). Medio siglo de Teoría de Respuesta a los ítems. *Anuario de Psicología. 52*, 41-66.
Muñiz, J, Hambleton, R. & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2); 115-135.
Muthén, B. & Lehman, J. (1985). Multiple group IRT modeling: Application to item bias analysis. *Journal of Educational Statistics. 10*(2), 133-142.

Navas Ara, M. J. (1996). Equiparación de puntuaciones. En J. Muñiz (Ed.), *Psicometría* (pp. 293-369). Madrid: Editorial Universitas, S. A.

Núñez Núñez, R. M., Hidalgo Montesinos, M. D., & López Pina, J. A. (2000). Influencia de la igualación iterativa en la detección del funcionamiento diferencial del ítem mediante medidas de área de Raju y el estadístico de Lord. *Psicothema. 12*(3), 495-502.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement. 25*(4), 373-383.

Park, D.-G. & Lautenschlager, G. (1990). Improving IRT Item Bias Detection with Iterative Linking and Ability Scale Purification. *Applied Psychological Measurement. 14*(2), 163-173.

Penfield, R. D. (2001). Assessing Differential Item Functioning among multiple groups: A comparison of three Mantel-Haenszel Procedures. *Applied Measurement in Education. 14*(3), 235-259.

Raju, N. (1988). The Area Between Two Item Characteristic Curves. *Psychometrika. 53*(4), 495-502.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhague: The Danish Institute for Educational Research.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. En W.J.van der Linder & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. En C.R.Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178-208). New York: Wiley.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika. 1*, 33-49.

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

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). A Monte Carlo Comparison of Seven Biased Item Detection Techniques. *Journal of Educational Measurement. 17*(1), 1-10.

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

Santor, D. & Ramsay, J. (1998). Progress in the Technology of Measurement: Applications of Item Response Models. *Psychological Assessment. 10*(4), 345-359.

Segal, D. O. (1983). Test characteristic curves, item bias and transformation to a common metric in IRT: A methodological artifact with serious consequences and a simple solution. Unpublished manuscript. Illinois: University of Illinois.

Shealy, R. & Stout, W. F. (1989, march). *A procedure to detect test bias present simultaneously in several items*. San Francisco: Paper presented at Annual Meeting of the American Educational Research Association.

Shealy, R. & Stout, W. F. (1993a). A Model Based Standardization Approach that Separates True Bias/DIF From Group Ability Differences and Detects Test Bias/DTF as well as Item Bias/DIF. *Psychometrika. 58*(2), 159-194.

Shealy, R. & Stout, W. F. (1993b). An item response theory model for test bias an differential test functioning. En P.W.Holland & H. Wainer (Eds.), *Differential item functioning*, Hilisdale, N. J.: LEA.

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

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of Approximation Techniques for Detecting Item Bias. *Journal of Educational Measurement. 22*(2), 77-105.

Stern, W. (1914). *The psychological methods of testing intelligence.* Baltimore: Warwick & York.

Stocking, M. & Lord, F. M. (1983). Developing a common metric in IRT. *Applied Psychological Measurement. 7*(2), 201-210.

Swaminathan, H. & Rogers, H. J. (1990). Detecting Differential Item Functioning using Logistic Regression Procedures. *Journal of Educational Measurement. 27*(4), 361-370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond Group-Mean Differences: The Concept of Item Bias. *Psychological Bulletin. 99*(1), 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use ot item response theory in the study of group differences in trace lines. En H.Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameter of item response models. En P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-113). Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika. 47*, 397-412.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika. 11*, 1-13.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning of Rasch models. *The Journal of Experimental Education. 72*(3), 221-261.

Wang, W.-C. & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement. 27*(6), 479-498.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee ability and ítem characteristic curve parameters. (Research Memorandum, 76-6. Princeton, NJ: Educational Testing Service.

Wright, B. D. & Mead, R. J. (1976). BICAL: Calibrating items with the Rasch model. (Research Memorandum, 23). Chicago: University of Chicago.

Manuscrito recibido en Enero de 2006
Aceptado para publicación en Mayo de 2006