Anuario de Psicología
1997, nº 74, 3-32

Evaluación del funcionamiento diferencial en items dicotómicos: una revisión metodológica

Juana Gómez
Universitat de Barcelona
Mª Dolores Hidalgo
Universidad de Murcia

Gran parte de la investigación psicométrica de las últimas décadas se centra en estudios del funcionamiento diferencial de los items (DIF) que analizan la imparcialidad de los tests con respecto a distintos grupos de sujetos. Este trabajo expone las principales técnicas de detección del DIF con items dicotómicos, valorando las ventajas e inconvenientes de cada una de ellas. La breve reseña de los primeros métodos incondicionales, en los que no se igualaban los grupos con respecto al nivel en el rasgo medido, da paso a una detallada evaluación de los métodos condicionales, clasificándolos en función de si las comparaciones entre grupos se llevan a cabo con respecto a una variable latente o a una variable observada. Se finaliza con un análisis de las aplicaciones prácticas de estas técnicas a diversos contenidos que ponen de manifiesto la necesidad de la utilización rutinaria de estudios de DIF, tanto en la evaluación de tests existentes como en el desarrollo de nuevos instrumentos de medida.
Palabras clave: *Funcionamiento diferencial de los items, técnicas de detección de DIF, aplicaciones de DIF.*

Most of the psychometric research reported in the last few decades is based on studies of differential item functioning (DIF), which analyze the impartiality of tests respect to different groups of subjects. This study describes the main techniques used in the detection of DIF with dichotomous items, and evaluates their advantages and disadvantages. The brief description of the first unconditional methods, in which groups were not matched with respect to the level of the trait measured, leads to a detailed evaluation of the conditional methods. These are classified according to whether comparisons between groups are based on a latent variable or an observed variable. The study concludes with an analysis of applications of

Dirección de las autoras: Juana Gómez. Facultat de Psicologia. Departament de Metodologia de les Ciències del Comportament, Pg. de la Vall d'Hebron, 171. 08035 Barcelona. e-mail:jgomez@psi.ub.es. Mª Dolores Hidalgo. Facultad de Psicología. Área de Metodología de las Ciencias del Comportamiento. Apdo. 4021. 30080 Murcia. e-mail: mdhidalg@fcu.um.es

these techniques to different contents, and emphasizes the need for the routine application of DIF studies both in the evaluation of existing tests and in the development of new tests.

Key words: *Differential item functioning, DIF detection techniques, empirical investigations of DIF.*

Los tests estandarizados se han convertido en una herramienta indispensable en la sociedad actual, tanto a nivel laboral como educativo. Sus resultados son utilizados ampliamente para medir aptitudes, habilidades y actitudes con el fin de tomar decisiones de selección, promoción, evaluación de la competencia, etc., que pueden afectar a la situación personal y laboral del individuo. Debido a esta masiva utilización de los tests para la toma de decisiones importantes, la cuestión de la presencia de un posible sesgo en los items que los componen se ha convertido en una preocupación central en la evaluación de la validez de los tests. En efecto, la existencia de sesgo puede representar una clara amenaza que atenta contra la validez de un test, en el que parte de sus items están beneficiando a ciertos grupos de la población en detrimento de otros de igual nivel en la característica medida.

Un instrumento de medición debe dar cuenta estrictamente de la variabilidad de los sujetos con respecto al atributo de interés; no debe estar afectado, en su función de medir, por las características del objeto de medida y, en el grado en que lo esté, la validez de dicho instrumento está seriamente limitada; para considerarse realmente válido, un instrumento de medición ha de garantizar resultados idénticos en sujetos que tienen el mismo nivel en el atributo medido, sea cual sea su grupo de pertenencia.' En este sentido, la relativa imparcialidad de los tests estandarizados ha sido en las últimas décadas campo fértil para numerosas investigaciones cuyo centro reside en la detección de items o tests que se comportan diferencialmente para grupos que difieren en lengua nativa, género, etnia, cultura, o cualquier otra variable que pueda constituir una fuente sistemática de variación ajena al rasgo medido por el test en cuestión.

Interesa comprobar si los tests son objetivos en su medición y las diferencias encontradas entre los sujetos con respecto a un determinado rasgo son reales, o bien son artefactos producto del instrumento con el que se ha medido dicho rasgo. Estos trabajos son popularmente conocidos como estudios de sesgo, aunque este término está siendo actualmente descartado por su carácter impreciso y confuso; en efecto, el término sesgo asume que el investigador examina tests o items que están sesgados mientras que lo único que puede inferirse realmente a partir de las respuestas de los sujetos es que hay diferencias en los resultados alcanzados por diferentes sujetos igualmente capaces que pueden deberse a distintas razones, una de las cuales es el sesgo. Un término más preciso, de uso cada vez más extendido, es el de funcionamiento diferencial de los items (*differential item functioning* o, abreviadamente, DIF), refiriéndose a que algunos items funcionan de modo distinto para sujetos o grupos que tienen habilidad similar. El término sesgo va más allá y pretende interpretar la causa por la que los items se comportan de modo diferente entre los grupos, situándose en un contexto de validez de constructo del instrumento de medida.

Shealy y Stout (1993a, 1993b) han insistido recientemente en la distinción entre ambos conceptos. Un test estará sesgado si en un grupo de sujetos tiene una validez menor que en otro grupo distinto y, por lo tanto, actúa injustamente con respecto al rasgo latente que pretende medir. Según estos autores, el sesgo se produce porque algunos de los items del test miden otros rasgos, además del que pretende medir el test, y los grupos pueden diferir en estos rasgos irrelevantes, aunque sean de un nivel semejante en el rasgo de interés. El DIF también se identifica cuando algunos de los items del test contribuyen a generar, en un nivel dado del atributo medido, diferencias entre grupos, pero no se presupone que el resto de los items cumpla el criterio de validez.

Los estudios de DIF se circunscriben a técnicas estadísticas que intentan determinar qué items funcionan diferencialmente para distintos grupos de sujetos, mientras que los estudios de sesgo añaden una explicación de porqué se producen estas diferencias. Es decir, los procedimientos estadísticos por sí solos no proporcionan evidencia de sesgo (Camilli y Shepard, 1994): si un estudio de DIF detecta los items de un test que funcionan diferencialmente entre dos o más grupos de sujetos, sólo el análisis lógico de estos items en el marco de la validez del test y en relación al constructo medido, identificará a los items realmente sesgados.

El presente trabajo se centra en las técnicas de detección de DIF, ya que el análisis del sesgo debe desarrollarse en cada aplicación concreta de un instrumento de medida en la que sea factible una medición sin equidad para distintos grupos de sujetos. Un ítem funcionará diferencialmente cuando dos grupos comparables de sujetos lo ejecuten de manera distinta; entendiéndose por comparables aquellos grupos de sujetos que, aunque distintos, poseen idéntico nivel con respecto al atributo medido por el test. Si el ítem en cuestión no presentase DIF, ambos grupos tendrían la misma probabilidad de éxito en la ejecución de dicho ítem; por el contrario, si el ítem está afectado por DIF habrá un diferencia en esta probabilidad que llevará a que uno de los dos grupos tenga una relativa ventaja con respecto al otro. En los estudios sobre DIF pueden compararse varios grupos de sujetos aunque generalmente se centran en dos: suele denominarse al grupo aventajado como grupo mayoritario o de referencia (R), mientras que el no aventajado o perjudicado se conoce como grupo minoritario o focal (F), comparable con el de referencia.

Es crucial el tener en cuenta la comparabilidad de los grupos, que lleva a remarcar la diferencia entre los términos impacto y DIF, para no confundirlos. Se entenderá que existe impacto de un ítem cuando las diferencias observadas en dos grupos distintos de la población en la ejecución del ítem en cuestión obedezcan a diferencias reales entre los grupos en la característica medida por el test; por el contrario, existirá DIF cuando sujetos con idéntico nivel en la característica medida –comparables– tengan distintas probabilidades de éxito para un determinado ítem, dependiendo del grupo al que pertenezcan.

El DIF puede aparecer de modo uniforme o no uniforme (Mellenbergh, 1982). Se denomina uniforme o consistente cuando no existe interacción entre el nivel en el atributo medido y la pertenencia a un determinado grupo; es decir, cuando la probabilidad de responder correctamente al ítem sea mayor para un

grupo que para el otro uniformemente a lo largo de todos los niveles del atributo. En cambio, existirá DIF no uniforme o inconsistente cuando se dé esta interacción, o sea, cuando la diferencia de las probabilidades de responder correctamente al ítem en los dos grupos no sea la misma a lo largo de todos los niveles del atributo.

Se han propuesto un gran número de técnicas, cada vez más sofisticadas, para detectar el DIF en sus diversas formas. Las primeras en utilizarse, denominadas aquí pioneras, estaban basadas en las diferencias en dificultad del ítem en cuestión; se las clasifica dentro de los métodos incondicionales en el sentido de que no se igualan los grupos con respecto al nivel del rasgo medido, por lo que actualmente se consideran descartadas ya que confunden el DIF con la discriminación del ítem. Por ello han sido reemplazadas por los métodos condicionales en los que, al emparejar los niveles del rasgo en los grupos, estos últimos pueden considerarse realmente comparables y permiten distinguir entre DIF e impacto.

A su vez, los métodos condicionales pueden diferenciarse en función de si las comparaciones entre grupos se llevan a cabo con respecto a una variable latente o a una variable observable (Millsap y Everson, 1993). Los primeros, denominados métodos de invarianza condicional no observada, especifican un modelo de medida y contrastan si los parámetros de dicho modelo permanecen invariantes para los diferentes grupos. Los segundos, etiquetados como métodos de invarianza condicional observada, utilizan la puntuación observada del test como un estimador del rasgo medido, sin especificación formal de modelo de medida, y verifican si las distribuciones de puntuaciones del ítem entre sujetos con valores iguales en la puntuación total del test son independientes del grupo de pertenencia.

En base a esta clasificación, y dada la importancia social del tema y su repercusión en la evaluación con cualquier instrumento de medida, el presente trabajo pretende ofrecer una exposición sucinta de las principales técnicas de detección de DIF, engarzándolas al final con las investigaciones empíricas que las han utilizado.

Métodos pioneros

Una de las primeras aproximaciones en el estudio del DIF fue el análisis de varianza (ANOVA) (Cleary y Hilton, 1968; Plake y Hoover, 1979). Bajo esta perspectiva un ítem funciona diferencialmente en los miembros de un grupo si las diferencias en términos absolutos entre la media de dicho grupo y las medias del resto de grupos sometidos a comparación son mayores que lo esperado en función del comportamiento de los otros items del test (Cleary y Hilton, 1968). Esta metodología, actualmente en desuso, resulta inapropiada para evaluar el DIF (Camilli y Shepard, 1987) ya que no detecta una buena parte de los items que realmente tienen DIF y ocasiona una elevada tasa de falsos positivos al no tener en cuenta la capacidad discriminativa del ítem, dejándose afectar por la dificultad media de los items y por el impacto entre grupos.

Como técnica complementaria al ANOVA, Angoff y Ford (1973) propusieron el análisis del delta-plot que se fundamenta en la transformación del índice de dificultad (TID). Esta técnica es esencialmente gráfica ya que el objetivo final es construir un diagrama de dispersión donde en el eje de ordenadas se representa el valor TID para cada ítem en el grupo focal y en el eje de abcisas el valor TID correspondiente al grupo de referencia. Los items que en cierto grado se alejen del eje principal del diagrama de puntos formado por la mayoría de items serían diagnosticados como items con DIF. Para identificar estos items no sólo se utiliza el gráfico sino que Angoff y Ford (1973) proponen el cálculo de un índice de distancia perpendicular de cada ítem al eje principal de la elipse: los valores más elevados de dicho índice, ya sean positivos o negativos, indican que el ítem funciona diferencialmente. Su sencillez de cálculo, unido a que no necesita tamaños muestrales elevados, hicieron de este procedimiento uno de los más utilizados durante los años 70; entre sus principales inconvenientes cabe destacar que, cuando los grupos presentan impacto, tiende a confundir éste con DIF (Shepard, Camilli y Averill, 1981) y que el cálculo del índice TID es dependiente de las características de las muestras donde ha sido obtenido, siendo por lo tanto inestable a través de distintas muestras. Aunque se han propuesto algunas variaciones (Angoff, 1982; Jensen, 1980; Rudner, Getson y Knight, 1980a,1980b; Shepard, Camilli y Williams, 1985) con la finalidad de mejorar la identificación de items con DIF en presencia de impacto entre grupos, en general estas alternativas no solucionan todas las deficiencias del método delta.

Green y Draper (1972, ver Shepard *et al.*, 1981) proponen evaluar el DIF representando gráficamente las correlaciones ítem-test obtenidas según la correlación biserial-puntual. Este procedimiento ha mostrado escaso acuerdo con otras técnicas de DIF (Ironson, 1982) y no es recomendable teniendo en cuenta que la correlación biserial-puntual es una medida de relación ítem-test afectada por la dificultad del ítem.

Métodos de Invarianza Condicional Observada

Estadísticos χ^2 tradicionales

En cualquier estudio de DIF el objetivo es evaluar si la probabilidad de responder correctamente a un mismo ítem en dos grupos es diferente; en otras palabras, si la proporción de aciertos es distinta a través de los grupos que están siendo comparados. Para conocer la significación de dichas diferencias es posible aplicar una prueba χ^2 tradicional de igualdad de proporciones. Siendo el nivel de habilidad una variable a controlar, las posibles diferencias en proporción de aciertos se evalúan condicionando sobre la habilidad. Son tres por tanto las variables que conformarían la tabla de contingencia a estudiar: la respuesta al ítem (R), el grupo de pertenencia (G) y el nivel de habilidad (H). Cuando la respuesta al ítem es dicotómica y se consideran dos grupos (R y F) la estructura de

los datos viene dada en la Tabla 1. En los estadísticos basados en χ^2, la tabla 1 de contingencia tridimensional se reduce a una tabla bidimensional al colapsar el nivel de habilidad. En esta subtabla bidimensional n_{1k} y n_{2k} son las frecuencias de sujetos que han contestado el ítem i en el intervalo k en los grupos R y F, respectivamente; $n_{.1k}$ y $n_{.2k}$ corresponden al número de sujetos que han acertado y que han fallado el ítem i en ambos grupos, respectivamente. n_{11k} y n_{21k} son las frecuencias relativas correspondientes al número de sujetos que han acertado el ítem en cada grupo (R y F), respectivamente, y n_{12k} y n_{22k} son las frecuencias relativas correspondientes al número de sujetos que han fallado el ítem también en cada grupo. Por último, $n_{..k}$ es el número total de sujetos que han contestado al ítem i en el intervalo k.

TABLA 1. TABLA DE CONTINGENCIA TRIDIMENSIONAL PARA EL ESTUDIO DEL DIF

		Respuesta al ítem		
Habilidad	*Grupo*	*Acierto*	*Fallo*	*Total*
H_1	Referencia	n_{111}	n_{121}	$n_{1.1}$
H_1	Focal	n_{211}	n_{221}	$n_{2.1}$
Subtotal 1		$n_{.11}$	$n_{.21}$	$n_{..1}$
H_2	Referencia	n_{112}	n_{122}	$n_{1.2}$
H_2	Focal	n_{212}	n_{222}	$n_{2.2}$
Subtotal 2		$n_{.12}$	$n_{.22}$	$n_{..2}$
⋮	⋮	⋮	⋮	⋮
H_k	Referencia	n_{11k}	n_{12k}	$n_{1.k}$
H_k	Focal	n_{21k}	n_{22k}	$n_{2.k}$
Subtotal k		$n_{.1k}$	$n_{.2k}$	$n_{..k}$
Total		$n_{.1.}$	$n_{.2.}$	$n_{...}$

Si se asume que la proporción de individuos que responden correctamente a un ítem en un intervalo k es una estimación de la probabilidad de acertar el ítem en dicho intervalo, entonces afirmar que un ítem no presenta DIF es probar la hipótesis de $p_{11k} = p_{21k}$ en los K intervalos, donde $p_{11k} = n_{11k} / n_{1k}$ y $p_{21k} = n_{21k} / n_{2k}$. Scheuneman (1979) propone comprobar dicha hipótesis mediante una prueba de χ^2 aplicada a los aciertos, con $(K-1) \times (r-1)$ grados de libertad, siendo r el número de grupos comparados.

Baker (1981) advierte sobre la utilización de este índice ya que, al considerar únicamente la proporción de aciertos, distorsiona los resultados, especialmente si existe impacto entre grupos, lleva a resultados diferentes en función de la equivalencia o no en los tamaños muestrales de los grupos F y R y es dudoso que siga realmente una distribución χ^2.

Ante las complicaciones planteadas al utilizar la prueba de Scheuneman, Camilli (1979, en Ironson, 1982) propone calcular la χ^2 de los aciertos y la de los errores y sumarlas; dicha suma se distribuye según una χ^2 con $K(r-1)$ grados de libertad. Esta prueba se ha denominado total, modificada o completa dado que para su cálculo se utilizan tanto los marginales de respuestas correctas como los de incorrectas.

Aunque la sencillez de las dos pruebas comentadas hace muy atractivo su uso no hay que olvidar que adolecen de ciertos inconvenientes: son sensibles a la presencia de impacto entre grupos, a diferentes tamaños muestrales, a frecuencias bajas o cercanas a cero y a cambios en los límites de los intervalos; además, no son capaces de evaluar el DIF no uniforme (Mellenbergh, 1982).

Procedimiento de Mantel-Haenszel

La prueba de Mantel-Haenszel (MH) fue desarrollada por Mantel y Haenszel (1959) y aplicada al análisis del DIF por Holland y Thayer (1988), considerándose una extensión de los enfoques tradicionales de χ^2 descritos anteriormente. El procedimiento de MH para detectar el DIF compara la ejecución de un ítem entre el grupo de referencia y el grupo focal a través de distintos niveles de un determinado criterio; se asume que en cada nivel los sujetos de uno y otro grupo son comparables y, si el ítem no presenta DIF, lo ejecutarían por igual. Raju, Bode y Larsen (1989) ponen de manifiesto que el número apropiado de niveles está en función del rango de puntuaciones y su estudio detecta como óptimo 4 en un rango de puntuaciones de 0 a 40.

Los datos se expresan en la forma de la tabla 1 calculándose un cociente de razones para cada tabla 2 × 2 (acierto o error en el ítem / pertenencia a R o F) de cada nivel del mencionado criterio, y un cociente de razones común –al que se denomina α– entre las diversas tablas 2 × 2 existentes. La estimación de α se expresa como:

$$\alpha = \frac{\sum_{k=1}^{K} n_{11k} n_{22k} / n_{..k}}{\sum_{k=1}^{K} n_{21k} n_{12k} / n_{..k}} \qquad (1)$$

pudiendo adoptar valores entre cero e infinito. La hipótesis nula de ausencia de DIF está representada por $\alpha = 1$; cuando el valor de alfa es superior a 1, el ítem favorece al grupo R, mientras que valores inferiores a 1 indican que el ítem es menos difícil para el grupo F. Dicha hipótesis nula se testa mediante el estadístico:

$$\text{MH} = \frac{(|\sum_{k=1}^{K} n_{11k} - \sum_{k=1}^{K} E(n_{11k})| - 0.5)^2}{\sum_{k=1}^{K} Var(n_{11k})} \qquad (2)$$

siendo $E(n_{11k}) = (n_{1.k})(n_{.1k}) / n_{..k})$ y $Var(n_{11k}) = (n_{1.k} n_{2.k} n_{.1k} n_{.2k}) / (n_{..k})^2 (n_{..k} - 1)$. Cuando la hipótesis nula es cierta, este estadístico sigue una distribución χ^2 con un grado de libertad. La ecuación incluye un factor de corrección por continuidad (0.5) que mejora la aproximación de la prueba MH a la distribución χ^2_1.

Si la estimación de alfa no es significativamente distinta de 1, al nivel de significación elegido, puede concluirse que el ítem analizado no tiene DIF, es decir, los grupos R y F ejecutan igual de bien el ítem cuando se controla su nivel en un determinado criterio relacionado con el test. En casi todos los estudios que utilizan el estadístico MH en el análisis del DIF se elige la puntuación total del test como criterio para establecer los niveles, dada la dificultad en encontrar criterios externos apropiados.

La vertiente negativa de la elección de la puntuación total del test como criterio se basa en que está contaminada por la inclusión de items con DIF. Para obviar este problema de circularidad, Holland y Thayer (1988) sugieren un procedimiento en dos etapas: en la primera se aplica la técnica de Mantel-Haenszel utilizando la puntuación total del test como criterio y se identifican los items con DIF; en una segunda fase, se eliminan del cálculo de la puntuación total todos los items que en la primera etapa han presentado DIF (con la excepción del ítem en estudio) y se repite de nuevo todo el proceso para cada uno de los items, presentaran o no DIF en la primera etapa. Este procedimiento de purificación ha sido calificado de muy eficaz en los estudios de Clauser, Mazor y Hambleton (1993) y de Navas y Gómez (1994), detectándose un gran incremento en la tasa de detecciones correctas. Fidalgo (1996) propone otro método de purificación iterativo en el que en cada iteración se eliminan los items con DIF y se recalculan de nuevo los estadísticos MH usando la puntuación total del sujeto en los items restantes como criterio de bloqueo, repitiendo la operación hasta que los items identificados con DIF sean los mismos en dos iteraciones sucesivas.

De todos los estadísticos descritos basados en χ^2, el procedimiento MH es el más potente (Holland y Thayer, 1988; López Pina, Hidalgo y Sánchez Meca, 1993) lo que, unido a su facilidad de cálculo y al hecho de no necesitar tamaños muestrales excesivamente grandes (Mazor, Clauser y Hambleton, 1992, hallan suficientes 200 sujetos por grupo), le ha convertido en uno de los métodos más utilizados para detectar items con DIF. Su inconveniente más grave reside en que no es eficaz en detectar DIF no uniforme (Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990; Uttaro y Millsap, 1994). Para paliar esta dificultad, Mazor, Clauser y Hambleton (1994) han propuesto una variante consistente en calcular de forma separada los estadísticos MH en el grupo de sujetos con puntuaciones altas en el test y en el grupo con puntuaciones bajas en el test, utilizando como punto de corte la media de la distribución de habilidad. Esta alternativa posibilita la utilización del procedimiento MH también en la detección del DIF no uniforme a un nivel de eficacia notable aunque, según el estudio de Fidalgo y Mellenbergh (1995), a costa de incrementar la tasa de error tipo I.

Procedimiento de Estandarización

El estadístico estandarizado propuesto por Dorans y Kulick (1986) pretende evaluar las diferencias entre las proporciones de acierto entre el grupo focal y el grupo de referencia en cada nivel de puntuación. Para tal fin se divide el continuo de habilidad en K intervalos y las respuestas de los sujetos a cada uno

de los items del test se organizan según la Tabla 1. En cada una de las K subtablas se calcula la diferencia entre la proporción de aciertos del grupo de referencia (p_{11k}) y la proporción de aciertos en el grupo focal (p_{21k}). El estadístico estandarizado viene dado por la siguiente expresión:

$$D_{EST} = \frac{\sum_{k=1}^{K} W_k[p_{21k} - p_{11k}]}{\sum_{k=1}^{K} W_k} \qquad (3)$$

donde W_k es un factor de ponderación que en un intervalo dado es el mismo para los dos grupos que se comparan. Aunque dicho factor de ponderación lo establece el propio investigador, Dorans y Kulick (1986) proponen los siguientes: el número de sujetos en el grupo total en el intervalo k, el número de sujetos en el grupo de referencia en el intervalo k, el número de sujetos en el grupo focal en el intervalo k, o el número de sujetos en un grupo estándar de referencia. Normalmente el factor de ponderación utilizado es el número de sujetos en el grupo focal.

D_{EST} es un índice cuyos valores varían entre –1 y +1. Valores positivos indican que el ítem favorece al grupo F y valores negativos indican que el ítem favorece al grupo R. En Dorans y Holland (1993) se encuentra una evaluación de los resultados obtenidos: si los valores de D_{EST} se encuentran entre –.05 y +.05, se considera que el ítem no presenta DIF; si los valores de D_{EST} se encuentran entre –.10 y –.05 o entre +.10 y +.05, es posible que el ítem presente DIF; por último, si los valores obtenidos son inferiores a –.10 o superiores a +.10 se debe realizar una revisión muy cuidadosa de esos items.

El procedimiento de estandarización también se aplica en el análisis del funcionamiento diferencial de distractores (FDD) en items de elección múltiple (Dorans, Schmitt y Bleistein, 1992) y en el estudio del DIF en items omitidos en tests de velocidad (Dorans y Holland, 1993).

Modelos loglineales y modelos logit

Como se ha señalado anteriormente, una de las desventajas de los procedimientos basados en K tablas bidimensionales es su incapacidad para evaluar DIF no uniforme. Los modelos loglineales y los modelos logit aportan una nueva perspectiva en el análisis de tablas de contingencia multidimensionales, resultando útiles para detectar los distintos tipos de DIF (Mellenbergh, 1982).

En el ajuste de este tipo de modelos se considera la tabla tridimensional 1. Una vez construida dicha tabla es posible formular distintas hipótesis acerca de cómo se distribuyen las frecuencias en cada una de las celdillas de la tabla. Dada la naturaleza categórica de las variables los modelos plausibles se expresan en términos multiplicativos, como productos de probabilidades marginales. Las dificultades generadas por el trabajo con modelos multiplicativos se solventan con una transformación de los datos con el fin de operar con modelos aditivos; en concreto, estos modelos son lineales en sus logaritmos naturales.

En el estudio del DIF, los modelos loglineales que interesa comparar para verificar cuál ofrece una mejor representación de la distribución de frecuencias observadas se concretan en:

Modelo 1 (DIF no uniforme): $ln\ E(n_{klp}) = \lambda+\lambda^{H}_{k}+\lambda^{G}_{l}+\lambda^{R}_{p}+\lambda^{HG}_{kl}+\lambda^{HR}_{kp}+\lambda^{GR}_{lp}+\lambda^{HGR}_{klp}$

Modelo 2 (DIF uniforme): $ln\ E(n_{klp}) = \lambda+\lambda^{H}_{k}+\lambda^{G}_{l}+\lambda^{R}_{p}+\lambda^{HG}_{kl}+\lambda^{HR}_{kp}+\lambda^{GR}_{lp}$

Modelo 3 (ausencia de DIF): $ln\ E(n_{klp}) = \lambda+\lambda^{H}_{k}+\lambda^{G}_{l}+\lambda^{R}_{p}+\lambda^{HG}_{kl}+\lambda^{HR}_{kp}$

Para probar el ajuste de cada uno de los modelos es posible utilizar varios estadísticos de bondad de ajuste, donde la idea básica es comparar las frecuencias esperadas, bajo el supuesto de que el modelo sea correcto, con las frecuencias observadas. Los más usuales son el estadístico χ^2 de Pearson y el estadístico G^2 o razón de verosimilitud (Bishop, Fienberg y Holland, 1975); estos estadísticos se distribuyen asintóticamente según una distribución χ^2 con los grados de libertad asociados al modelo ajustado.

Dado que los modelos son anidados, puede compararse su ajuste y eliminar los términos que no tengan un efecto significativo hasta conseguir la mejor explicación del comportamiento de los datos. En función de ello, se concluirá que el ítem presenta DIF no uniforme si el modelo 1 es el que ofrece mejor ajuste, ya que indica que la respuesta al ítem depende de los términos de interacción GxR y HxGxR; si este último término puede eliminarse y, por lo tanto, el modelo 2 constituye una mejor representación de la realidad, el ítem presentará DIF uniforme; por último, si ninguno de los dos términos de interacción se muestran necesarios, el modelo 3 de ausencia de DIF será el seleccionado.

Si se considera la variable respuesta al ítem como una variable dependiente y se realiza una transformación logit del tipo ln $[p / 1 - p]$, siendo p la probabilidad de éxito, entonces cualquier hipótesis que se someta a prueba se verificaría al ajustar modelos logit. Así, los anteriores modelos loglineales pueden formularse en términos logit:

Modelo 1 (DIF no uniforme): $ln\ (p_{kil} / 1 - p_{kil2}) = \eta+\eta^{H}_{k}+\eta^{G}_{l}+\eta^{HG}_{kl}$

Modelo 2 (DIF uniforme): $ln\ (p_{kil} / 1 - p_{kil2}) = \eta+\eta^{H}_{k}+\eta^{G}_{l}$

Modelo 3 (ausencia de DIF): $ln\ (p_{kil} / 1 - p_{kil2}) = \eta+\eta^{H}_{k}$

Los parámetros en la formulación logit pueden ser interpretados similarmente a los parámetros en un modelo de ANOVA, donde η es el efecto total de la dificultad del ítem, η^{G}_{l} es el efecto principal de la variable grupo, η^{H}_{k} es el efecto principal de la variable habilidad de los sujetos (agrupada en intervalos) y η^{HG}_{kl} es el efecto de la interacción habilidad x grupo. La comparación entre los modelos con respecto a su ajuste para seleccionar el más adecuado a los datos sigue las pautas indicadas para los modelos loglineales.

En la aplicación de los modelos logit y loglineales se utilizan todos los items del test (incluidos los que presentan DIF) en la obtención de las puntuaciones totales en el mismo, lo que puede alterar los resultados. Van der Flier, Mellenbergh, Adèr y Wijn (1984) proponen un método logit iterativo de purificación de la habilidad, con la finalidad de mejorar la detección de items que presentan DIF. Este método logit ha resultado más preciso en la identificación de items con DIF que el procedimiento logit no iterativo (Fidalgo y Paz, 1995; Kok, Mellenbergh y Van der Flier, 1985; Van der Flier *et al.*, 1984). Sin embargo, Mellenbergh (1989) señala que el método logit iterativo puede ser impreciso en la

evaluación del DIF cuando el porcentaje de items con DIF en el test es grande, en este caso los items sin DIF pueden ser insuficientes para obtener una buena aproximación de la habilidad de los sujetos.

Regresión logística

La utilización de la regresión logística (RL) como procedimiento para el análisis del DIF fue sugerida por Swaminathan y Rogers (1990) como una alternativa a los métodos basados en la teoría de respuesta al ítem (TRI) para detectar tanto DIF uniforme como no uniforme.

Los modelos de RL son idénticos a los modelos de regresión lineal con la diferencia de que la variable criterio no es de naturaleza cuantitativa sino categórica (normalmente dicotómica). En cada ecuación de regresión, uno de los items figura como variable dependiente, mientras que el nivel de habilidad del sujeto, la pertenencia a los grupos R o F y la interacción entre ambos constituyen las variables independientes. La estimación de habilidad que se usa generalmente es la puntuación total del test, tratada como variable continua, aunque este modelo tiene gran flexibilidad y puede incluir otras estimaciones de habilidad, variables concomitantes o alguna combinación.

La ecuación general para un modelo de regresión logística es:

$$P(x = 1) = e^z / (1 + e^z) \tag{4}$$

donde,

$$z = \tau_0 + \tau_1 H + \tau_2 G + \tau_3 (HG) \tag{5}$$

siendo x la respuesta al ítem, H el nivel de habilidad del sujeto, G el grupo de pertenencia (R o F) y $H \times G$ el producto de dos variables individuales, H y G. El parámetro τ_2 corresponde a las diferencias de grupo en la ejecución del ítem, mientras que el parámetro τ_3 corresponde a la interacción entre grupo y nivel de habilidad. Un ítem muestra DIF uniforme si τ_2 es distinto de cero y τ_3 es cero, y DIF no uniforme si τ_3 es distinto de cero, sea o no τ_2 igual a cero. La hipótesis nula a comprobar es si ambos parámetros no son significativamente distintos de cero.

Los coeficientes de RL se estiman por el método de máxima verosimilitud. Una vez que un modelo ha sido estimado debe comprobarse su adecuación o la relevancia de sus variables componentes. La significación estadística del efecto de las variables explicativas sobre la variable dependiente puede evaluarse mediante el test de Wald o el de razón de verosimilitud. En el primer caso, la hipótesis nula de que un determinado parámetro logístico τ_p es igual a cero se evalúa mediante el test de Wald que tiene una distribución χ^2 y, para variables con un grado de libertad, equivale al cuadrado de la razón entre el parámetro estimado y su error estándar. En el segundo caso, se evalúan los cambios en el ajuste del modelo a los datos cuando se elimina alguna variable. Para ello, se comparan los siguientes modelos:

Modelo 1 (DIF no uniforme) $z = \tau_0 + \tau_1 H + \tau_2 G + \tau_3 (HG)$

Modelo 2 (DIF uniforme) $z = \tau_0 + \tau_1 H + \tau_2 G$
Modelo 3 (ausencia de DIF) $z = \tau_0 + \tau_1 H$

La comparación del modelo 1 versus el modelo 2 comprueba si el término de interacción es necesario; la diferencia entre ambos modelos se evalúa mediante el estadístico de razón de verosimilitud, con un grado de libertad correspondiente al coeficiente τ_3. Si el término de interacción no demuestra su significación, se comparan los modelos 2 versus 3 (con otro grado de libertad correspondiente al coeficiente τ_2) para comprobar si la pertenencia al grupo tiene un efecto relevante sobre la respuesta al ítem. En definitiva, se escoge el modelo que describe los datos con los mínimos términos.

El procedimiento RL es en teoría superior a otros métodos basados en tablas de contingencia ya que tiene en cuenta la naturaleza continua de la escala de habilidad y es capaz de detectar todo tipo de DIF. Comparando con MH, RL puede calificarse de más general y flexible, aunque también más costoso computacionalmente. De hecho, MH puede considerarse un modelo de RL donde la habilidad se ha categorizado y no se formula término de interacción entre habilidad y grupo. La utilización del término de interacción entre habilidad y grupo ya fue propuesta por Mellenbergh (1982) mediante un modelo loglineal y de hecho el modelo de RL equivale a un modelo logit cuando las variables independientes se codifican como variables dummy. Sin embargo, el procedimiento de RL es preferible al de loglineal ya que trata la habilidad como una variable continua, utilizando toda la información disponible.

Su principal inconveniente radica en la escasa eficacia para detectar DIF estrictamente uniforme; según Swaminathan y Rogers (1990), el término de interacción puede afectar adversamente a la potencia del procedimiento cuando el DIF presente es sólo uniforme ya que se pierde innecesariamente un grado de libertad. Para mejorar la tasa de detección en caso de DIF uniforme, se han propuesto dos tipos de procedimientos de purificación. Navas y Gómez (1994) proponen un procedimiento en dos etapas similar al aconsejado por Holland y Thayer (1988) para el estadístico MH: en una primera etapa, se analizan todos los items mediante RL, utilizando como criterio la puntuación total del test, y en una segunda etapa se vuelve a aplicar RL para todos los items pero con el criterio de habilidad purificado mediante la eliminación de todos los items que han presentado DIF en el paso anterior, excepto el ítem bajo estudio; la utilización de este procedimiento frente a RL sin purificar el criterio mejora la tasa de identificación de items con DIF, disminuyendo además la tasa de falsos positivos (Navas y Gómez, 1994). Posteriormente, Gómez y Navas (1996) han propuesto utilizar RL con un procedimiento iterativo en el que la purificación de la habilidad se realiza paso a paso; dicho procedimiento, similar al método logit iterativo propuesto por Van der Flier *et al.* (1984), mejora la eficacia en la evaluación del DIF uniforme en relación al método de dos etapas en el que la purificación se realiza simultáneamente.

SIBTEST

Este procedimiento, desarrollado por Shealy y Stout (1993a, 1993b) y fundamentado en un modelo de TRI multidimensional no parámetrico presenta

características tanto de los métodos de invarianza condicional observada como de los métodos de invarianza condicional no observada (Millsap y Everson, 1993), por lo que realmente cabría ubicarlo en cualquiera de ambas clasificaciones, tanto como en ninguna de ellas.

SIBTEST posibilita detectar el DIF para un ítem aislado o para un grupo de items de modo simultáneo. Esta peculiaridad da pie al estudio de dos fenómenos de DIF: amplificación y cancelación. El efecto de amplificación del DIF se produce cuando un subconjunto de items presenta individualmente DIF sobre un grupo dado, bajo estas circunstancias SIBTEST permite estudiar el efecto conjunto de dichos items sobre dicho grupo. Por otro lado, el fenómeno de cancelación del DIF se produce cuando un subconjunto de items de un test presenta DIF sobre el grupo R y otro subconjunto lo presenta sobre el grupo F; esta situación puede provocar que en el test el efecto del primer subconjunto de items cancele el efecto del segundo subconjunto.

Shealy y Stout distinguen entre habilidad objetivo y determinantes ruido. Por habilidad objetivo se entiende aquel rasgo (θ) que el test pretende medir, y por determinantes ruido (η) una o más habilidades o constructos que no se intentan medir con el test pero que afectan a uno o más items del test; esto implica que el DIF estaría provocado por las diferencias entre grupos en esa(s) habilidad(es) ruido, debido a que los items implicados serían multidimensionales.

La evaluación del DIF en un conjunto de n items asume que del total de éstos una parte de ellos presentan DIF, siendo estos items el objeto de estudio. Para aplicar el procedimiento propiamente dicho, Shealy y Stout (1993b) definen para un test dos puntuaciones totales: una (X) compuesta por aquellos items (n_x) que formarían el subtest válido porque miden solamente la θ objetivo, y otra (Y) formada por la suma de las respuestas dadas a los items que constituirían el subtest bajo estudio, es decir, por aquel ítem o aquellos items (n_y) que miden no sólo la θ objetivo sino también otras habilidades distractoras. Una vez obtenidas estas puntuaciones, los sujetos del grupo focal y del grupo de referencia se agrupan en función de las puntuaciones totales X obtenidas en el subtest válido y se comparan con respecto a sus resultados en el subtest estudiado.

El trabajo de Nandakumar (1993) constata la efectividad de SIBTEST en la detección de amplificación y cancelación del DIF en diferentes tamaños muestrales y diferentes tamaños de test. Nandakumar (1993), Narayanan y Swaminathan (1994) y Shealy y Stout (1993b) muestran como, desde la perspectiva del análisis del ítem, los resultados obtenidos por MH y SIBTEST concuerdan en relación a la dirección y cantidad del DIF estimado. Hasta el momento SIBTEST había encontrado dificultades en la evaluación del DIF no uniforme, sin embargo trabajos recientes (Li y Stout, en prensa) han extendido este procedimiento a la detección del DIF no uniforme.

Métodos de Invarianza Condicional No observada

La mayor parte de los métodos que se pueden incluir en este apartado asumen como modelo de medida un modelo de TRI. En un modelo de TRI dicotó-

mico unidimensional, la curva característica del ítem (CCI) puede estar definida por el modelo logístico de un parámetro (1-p), el modelo logístico de dos parámetros (2–p) o el modelo logístico de tres parámetros (3–p). En el modelo de 3–p, la probabilidad de obtener una respuesta correcta a un ítem i en un nivel de habilidad dado θ vendría dada por (Birnbaum, 1968):

$$P_i(\theta) = c_i + (1 - c_i)\{1 + exp[-1.702a_i(\theta - b_i)]\}^{-1} \quad (6)$$

donde b_i se define como un parámetro de localización que determina la posición de la curva a lo largo de la escala de θ y es un indicador de la dificultad del ítem, a_i, también conocido por parámetro de discriminación, representa la inclinación de la CCI en el punto de inflexión de la misma, y c_i es el parámetro de pseudoazar. Si se asume que el ítem no puede ser respondido correctamente por mero azar, es decir, que $c_i = 0$, entonces el modelo sería el de 2–p (Birnbaum, 1968). Si además se asume que los items tienen todos la misma capacidad discriminativa, es decir, que a es un valor constante en todos los items de un test, entonces el modelo sería el de 1–p o modelo de Rasch (Rasch, 1960).

Un ítem no presenta DIF si las CCIs del grupo focal y del grupo de referencia coinciden, es decir, si se cumple que $P_{iR}(\theta) = P_{iF}(\theta)$. La igualdad de las CCIs en los distintos grupos implica que los parámetros que las definen sean los mismos: un ítem no presenta DIF si $a_{iR} = a_{iF}$, $b_{iR} = b_{iF}$ y $c_{iR} = c_{iF}$. Por el contrario, un ítem presenta DIF si se da una de las tres condiciones siguientes: 1) $a_{iR} = a_{iF}$ y $b_{iR} \neq b_{iF}$ (DIF uniforme), 2) $a_{iR} \neq a_{iF}$ y $b_{iR} = b_{iF}$ (DIF no uniforme), y 3) $a_{iR} \neq a_{iF}$ y $b_{iR} \neq b_{iF}$ (DIF no uniforme mixto).

Desde la perspectiva de la TRI, un paso previo a la aplicación de cualquier medida de evaluación del DIF conlleva disponer de las estimaciones de los parámetros de los items. La precisión en la estimación de parámetros puede provocar diferencias en la identificación de DIF, por lo que un punto clave es la elección del procedimiento de estimación más apropiado (Cohen, Kim y Subkoviak, 1991). Uno de los procedimientos más utilizado ha sido el de máxima verosimilitud conjunta (MVC), aunque cuando se trabaja con tamaños muestrales inferiores a 500 sujetos y longitud del test por debajo de 20 items es más apropiado utilizar el método de máxima verosimilitud marginal (MVM) (Cohen, Kim y Subkoviak, 1991; Kim, Cohen y Kim, 1994; McLaughlin y Drasgow, 1987).

Una de las propiedades de la TRI es la de invarianza de los parámetros, bien sean parámetros de items o bien parámetros de θ. Cuando se trabaja con estimaciones, este supuesto se suaviza e implica que los parámetros estimados deben, dentro de ciertos límites, ser los mismos en cada muestra extraída de una misma población o al menos estar linealmente relacionados. En este sentido, cuando se realiza una comparación entre parámetros estimados en distintos grupos es necesario igualarlos, es decir, encontrar las constantes que permitan transformar los parámetros de una muestra a la misma escala que los parámetros de la otra. Se dispone de distintos métodos de obtención de las constantes de igualación (Holland y Rubin, 1982), siendo los más utilizados los basados en las curvas características.

El principal problema cuando se identifican items con DIF radica en la inclusión en el proceso de igualación de parámetros de todos los items del test, es decir, tanto los items con DIF como sin DIF. La incorporación de los items con DIF en la igualación afecta directamente al cálculo de las constantes de igualación falseando la escala de transformación de los parámetros y provocando errores en la identificación del DIF (Candell y Drasgow, 1988; Lord, 1980; Miller y Oshima, 1992; Park y Lautenschlager, 1990). La solución a esta problemática conlleva procedimientos de purificación del test eliminando del cálculo de las constantes de igualación aquellos items en los que en un primer paso se ha detectado presencia de DIF, y volviendo a evaluarlos con las nuevas constantes de igualación.

Lord (1980) propone el siguiente procedimiento de purificación del test: 1) estimar los parámetros de los items conjuntamente para ambos grupos, estandarizando (media = 0 y varianza = 1) sobre el parámetro de dificultad y no sobre θ, 2) fijar los parámetros de pseudoazar a los valores que se han obtenido en el paso anterior y a continuación estimar b_i y a_i para cada grupo por separado, estandarizando sobre b, 3) identificar los items con DIF, 4) eliminar del test los items identificados con DIF en el paso anterior, 5) estimar θ para los sujetos de ambos grupos, 6) estimar los parámetros de los items del test fijando θ a los valores que han sido obtenidos en el paso anterior y 7) volver al paso 4.

Candell y Drasgow (1988) proponen que la purificación de los items se realice en el proceso de igualación, y que sea ésta la que se implemente iterativamente. La ventaja de este procedimiento frente al de Lord es su facilidad de implementación dado que no requiere la reestimación iterativa de los parámetros de los items y de θ.

Park y Lautenschlager (1990) proponen otra modificación del procedimiento de Lord que denominaron con las siglas M-LTP: consiste en la estimación iterativa de los parámetros de los items eliminando iterativamente de las estimaciones de habilidad el efecto de los items identificados con DIF. Este procedimiento de purificación de la habilidad mostró peores resultados en la identificación del DIF que el propuesto por Candell y Drasgow (Park y Lautenschlager, 1990). Por ello, Park y Lautenschlager (1990) y Lautenschlager, Flaherty y Park (1994) proponen un método de igualación iterativa y purificación de la habilidad (ILAP) que constituye una combinación de los dos procedimientos anteriores; dichos autores concluyen que los procedimientos iterativos son más eficaces en la evaluación del DIF que los no iterativos, y que el método ILAP es más preciso que el procedimiento de igualación iterativa de Candell y Drasgow (1988) y que el método M-LTP de Park y Lautenschlager (1990). Miller y Oshima (1992) proponen un procedimiento en dos etapas que presentó mejores resultados cuando el número de items con DIF en el test fue elevado (20% o más) y la magnitud del DIF fue moderada (una diferencia en b de .35).

En general, la evidencia experimental sugiere que los procedimientos iterativos pueden mejorar la precisión de la detección del DIF en comparación al procedimiento no iterativo y la cuestión estaría en si es más efectivo un procedimiento de purificación en dos etapas o un procedimiento iterativo.

Medidas de Área

Dado que la TRI considera que un ítem no presenta DIF cuando los parámetros estimados en cada grupo son iguales, la representación gráfica de las CCIs de los grupos F y R debería coincidir; si difiere, esta diferencia sería un indicador de DIF. De este modo, uno de los procedimientos propuestos para detectar el DIF ha sido el cálculo de la diferencia entre la CCI del grupo R y la CCI del grupo F, es decir, obtener el área entre ambas funciones.

Rudner (1977; Rudner *et al.*, 1980a,1980b) propone una medida de área sin signo basada en la comparación de las CCIs teóricas según los parámetros estimados en cada grupo por separado que vendría dada por:

$$A_i = \sum_{\theta=-4.00}^{\theta=+4.00} | P_{iR}(\theta) - P_{iF}(\theta) | \Delta_\theta \qquad (7)$$

La probabilidad de acertar el ítem en cada grupo se calculará en cada uno de los valores de θ de –4 a +4 en incrementos $\Delta_\theta = .005$ y utilizando la ecuación correspondiente según el modelo de TRI que se ajuste a los datos (1–*p*, 2–*p* o 3–*p*). Si el valor de A_i es elevado, entonces el ítem funciona diferencialmente; por el contrario, si $A_i = 0$, el ítem en cuestión no presenta DIF. Ironson y Subkoviak (1979) proponen una medida de área con signo basada en el procedimiento de Rudner (1977), que resulta ineficaz si el DIF es no uniforme; en estas situaciones el valor de la misma puede ser muy pequeño o incluso cero por el efecto de compensación de las diferencias de ambas CCI.

Linn, Levine, Hastings y Wardrop (1981) proponen cuatro índices de DIF: *a*) *Área base superior* que es el área entre la CCI del grupo R y la CCI del grupo F, rango de (–3, +3), en aquellos intervalos en los que la CCI del grupo R está por encima de la CCI del grupo F, *b*) *Área base inferior* sería el área entre la CCI del grupo de R y la CCI del grupo F en aquellos intervalos en los que la CCI del grupo F está por encima de la CCI del grupo R, *c*) *Área total* que se obtiene sumando las dos medidas de área anteriores y *d*) *RDMC* o raíz cuadrada de la diferencia media cuadrática. Los dos primeros índices proporcionan información sobre la dirección del DIF, mientras que los últimos son indicadores de la cantidad de DIF. Se interpretan del mismo modo que la medida de área propuesta por Rudner.

Raju (1988, 1990) desarrolla un conjunto de medidas de área exactas (con signo y sin signo) que, en relación a las medidas anteriores, permite probar su significación a través de una prueba *Z*. La expresión general de estas medidas, basadas en la integración continua, viene dada por:

$$A_i = \int_{-\infty}^{\infty} f\,[P_R(\theta) - P_F(\theta)]\, d(\theta) \qquad (8)$$

La función *f* puede especificarse con signo o sin signo. La expresión general de la ecuación 8 adopta distintas formas según el modelo de TRI con el que se esté trabajando y si se cumplen o no ciertas condiciones en los parámetros de los

ítems, como igualdad de los parámetros de discriminación y/o de los parámetros de pseudoazar.

Las medidas de área propuestas por Kim y Cohen (1991) calculan el área entre dos CCIs integrando sobre la diferencia entre las funciones de respuesta a un mismo ítem de dos grupos en un intervalo cerrado entre dos puntos finitos θ_1 y θ_2.

Los resultados obtenidos al usar una medida de área pueden estar sujetos a fluctuaciones muestrales; así, sin conocer el error típico asociado, no es posible determinar con un grado de confianza si dos CCIs difieren realmente o solamente por error de muestreo. La ventaja fundamental de las medidas propuestas por Raju (1988) es que van acompañadas de una prueba de significación que permite probar, a un nivel de confianza establecido, si el área entre dos CCIs es significativamente diferente de cero (Raju, 1990). En general, las medidas exactas de área se prefieren a las medidas basadas en intervalo pero, cuando se trabaja con tamaños muestrales pequeños, puede darse el caso de que en alguna región de la escala de habilidad las frecuencias sean bajas, hecho que afecta a la medida de área exacta mucho más que a la medidas de área basadas en intervalo; para estas últimas, es posible seleccionar el intervalo en el que se van a calcular las diferencias entre CCIs y así controlar la influencia de frecuencias bajas en esos niveles de θ. Una solución a esta deficiencia de las medidas exactas de área sería definir la función de densidad de θ de tal modo que ponderara con mayor peso el intervalo de habilidad de interés (Raju, 1988). Otra limitación importante de las medidas exactas de área es que no es posible calcularlas cuando el test se ajusta al modelo de 3–p y los parámetros de pseudoazar son diferentes en los grupos comparados, ya que en estas situaciones el valor del área es infinito; las medidas de área propuestas por Kim y Cohen (1991) pueden ser una alternativa válida teniendo en cuenta que en el resto de modelos dicotómicos aportan resultados muy similares a los obtenidos por las medidas exactas de área (Kim y Cohen, 1991).

Comparación de parámetros

En la evaluación del DIF no sólo se pueden utilizar procedimientos basados en comparar las CCIs, sino que también es factible comparar directamente los parámetros estimados de los ítems en dos o más grupos. Lord (1980) propone, cuando los parámetros son estimados por MVC o MVM, la siguiente prueba estadística:

$$\chi^2 = V'S^{-1}V \tag{9}$$

donde V es el vector de diferencias entre los parámetros estimados para un ítem en el grupo de referencia y los parámetros estimados para ese mismo ítem en el grupo focal y S^{-1} es la inversa de la matriz de varianza-covarianza asintótica para los vectores de diferencias entre parámetros.

El estadístico propuesto por Lord, bajo la hipótesis nula, sigue una distribución χ^2 con tantos grados de libertad como contrastes se hayan establecido (como columnas del vector V). Si se trabaja bajo el modelo de 3–p, Lord (1980) reco-

mienda fijar los parámetros *c* a un valor determinado previamente, dado que cuando se estima mediante MVC puede ser indeterminado. Para resolver este problema, Lord (1980) propone estimar conjuntamente (para ambos grupos) los parámetros de los items, estandarizando sobre el parámetro de dificultad y no sobre la θ, fijar los parámetros de pseudoazar a los valores obtenidos y a continuación estimar los parámetros de dificultad y discriminación para cada grupo por separado.

El estadístico de Lord se calcula fácilmente y proporciona una prueba de significación, sin embargo presenta algunas deficiencias. En primer lugar, la estimación de la θ se realiza sobre todos los items del test, de tal modo que se incluyen también los items con DIF; utilizar todos los items en la estimación de la θ contamina los resultados en cuanto a la identificación correcta de items con DIF y puede provocar el incumplimiento de los supuestos de unidimensionalidad e independencia local de los items. En segundo lugar, se asume que θ es conocida, sin embargo los valores de θ son estimados. El uso de parámetros estimados tiende a inflar la tasa de error tipo I, siendo este efecto menor cuando el tamaño muestral es grande, dado que al estimar los parámetros la utilización de *S* como estimador de la matriz de varianza-covarianza puede ser inadecuada. McLaughlin y Drasgow (1987) demostraron en un estudio de simulación que bajo los modelos de 2-*p* y 3-*p* los errores típicos de los parámetros estimados por MVC estaban negativamente sesgados. En tercer lugar, se desconoce el funcionamiento de la prueba cuando los tamaños muestrales son pequeños. Tampoco se conoce el tamaño muestral mínimo para que este estadístico converja a una distribución χ^2. Por último, puede rechazar la hipótesis nula cuando otras medidas de DIF como la medida de área sin signo indica que el área entre las dos CCIs es muy pequeña (Linn *et al.*, 1981); esto se hace especialmente evidente en presencia de impacto entre grupos, en tamaños muestrales grandes y cuando el porcentaje de items con DIF en el test es alto (40%) (Hidalgo, 1995; Hidalgo y López Pina, 1997). Millsap y Everson (1993) recomiendan que cuando el estadístico χ^2 de Lord sea significativo se combine con alguna medida de área para comprobar el DIF del ítem. Hidalgo y López Pina (1997) apuntan a que otros métodos de invarianza condicional observada, tales como RL también pueden ser efectivos como complemento al estadístico de Lord.

Comparación de modelos

Thissen, Steinberg y Gerrard (1986) proponen un procedimiento que comprueba si los parámetros de las CCIs de un mismo ítem administrado a grupos distintos difieren. Se compara un modelo denominado aumentado (A), en el que se especifica que los parámetros de los items no son los mismos para los dos grupos, con un modelo denominado compacto (C) que establece la restricción de igualdad de los parámetros de los items en los dos grupos. Teniendo en cuenta que el modelo C se encuentra anidado en el modelo aumentado, se calcula G^2 en ambos modelos y se somete a prueba la hipótesis de no diferencias entre los parámetros comparando ambos modelos a través del estadístico $G^2 = G^2[C] - G^2[A]$, que bajo la hipótesis nula sigue una distribución χ^2 con grados de libertad igual

a la diferencia entre el número de parámetros de items estimados en el modelo A y en el modelo C. Si el valor obtenido es mayor que el valor teórico de la distribución χ^2, se rechaza la hipótesis nula y por lo tanto el modelo C, concluyendo que el ítem o items especificados en el estudio presentan DIF.

Thissen, Steinberg y Wainer (1988, 1993) señalan que el estadístico χ^2 de Lord y el estadístico G^2 son asintóticamente equivalentes, la diferencia fundamental entre ellos se refiere a que el cálculo del estadístico G^2 no requiere la matriz de covarianzas entre parámetros estimados, mientras que χ^2 sí. Las limitaciones principales del procedimiento de Thissen *et al.* (1988) residen en su incapacidad para detectar diferencias entre parámetros menores de 0.3 y en su dependencia del tamaño muestral. Como ventaja, puede citarse que si se trabaja con pocos items y la tabla de contingencia (sujetos × patrones de respuesta) contiene pocas celdillas con cero, es posible evaluar el funcionamiento diferencial de todo el test (DTF) o de un subconjunto de items de modo simultáneo; en estos casos, el rechazo del modelo compacto conllevaría el posterior análisis de cada ítem por separado para identificar los que presentan DIF.

Otra aproximación la constituye el trabajo de Kelderman (1989) basado en la utilización del modelo de Rasch desde la perspectiva de un modelo loglineal. La hipótesis nula de partida se establece en los mismos términos que el procedimiento anterior, sin embargo la diferencia fundamental entre ambos radica en que bajo el procedimiento loglineal los contrastes entre parámetros se definen directamente en el modelo. Además, es más fácil la formulación de distintas hipótesis acerca del DIF. Por el contrario, parece que solamente es posible representar como un modelo loglineal la familia de modelos de Rasch, limitando el campo de aplicación. Kelderman y McReady (1990) extendieron el uso del modelo loglineal de Rasch para la evaluación del DIF, a otras situaciones tales como la inclusión de una dimensión latente categórica, aplicando modelos de clase latente. Aunque esta perspectiva ofrece muchas posibilidades de estudio del DIF, por ejemplo en el análisis de distractores en los items de elección múltiple (Westers y Kelderman, 1991), sin embargo, adolece de problemas muy serios. En primer lugar, conforme aumenta el número de items del test, aumenta el número de celdillas de la tabla de contingencia provocando que no sea posible su análisis, por lo que el investigador tiene que seleccionar cada vez un subconjunto de items y analizarlos por separado. En segundo lugar, y en cuanto a la estimación de parámetros, el algoritmo puede necesitar un número bastante grande de iteraciones para alcanzar la convergencia y ésta no siempre se consigue, especialmente cuando el modelo a ajustar es complejo o las soluciones de partida del algoritmo no son idóneas (Kelderman y McReady, 1990).

En general, los procedimientos de comparación de modelos presentan como una seria deficiencia su complejidad de cálculo y el costo computacional.

Valoración de procedimientos y aplicaciones prácticas

A la vista de las técnicas expuestas, cabe preguntarse cuál se debe utilizar en un estudio de DIF. La respuesta no es sencilla pues depende, entre otras, de va-

riables tales como el tamaño muestral de los grupos en los que se está estudiando el DIF, el tipo de DIF (uniforme o no uniforme) y el valor de dificultad y discriminación de los items. La selección de una u otra técnica de DIF requiere considerar estas y otras cuestiones para cada estudio específico de DIF.

En primer lugar, es imprescindible utilizar técnicas donde los grupos se igualen en el rasgo o habilidad medido por el test, ya que en principio es una garantía para no confundir impacto con DIF; en este sentido, los métodos pioneros tales como el ANOVA y el método delta-plot no son aconsejables en el estudio del DIF, aunque su sencillez los haga muy atractivos. En segundo lugar, las técnicas que permiten probar la ausencia de DIF a un cierto nivel de confianza, es decir, las aproximaciones que disponen de una prueba de significación, resultan más útiles a efectos de tomar decisiones más objetivas y fiables, con la ventaja de hacerlas comparables a las de otros trabajos. En tercer lugar, cualquiera que sea la técnica utilizada, es conveniente utilizar algún procedimiento de purificación de la medida de habilidad, ya que reduce el número de falsos positivos e incrementa el número de identificaciones correctas (Candell y Drasgow, 1988; Clauser *et al.*, 1993; Gómez y Navas, 1996; Holland y Thayer, 1988; Kok *et al.*, 1985; Navas y Gómez, 1994; Park y Lautenschlager, 1990; Van der Flier *et al.*, 1984).

A las pautas generales anteriores cabe añadir que, si los tamaños muestrales de los grupos a comparar son amplios (a partir de 1000 sujetos), es preferible utilizar técnicas de detección de DIF derivadas de la TRI que implican una estimación de la habilidad latente. Las técnicas que utilizan la puntuación observada en el test como criterio de equiparación, asumiendo que esta puntuación es una estimación de la habilidad latente del sujeto, pueden resultar imprecisas en la detección de DIF principalmente cuando el test contiene items de distinta discriminación, es decir, no se ajusta al modelo de Rasch. El procedimiento propuesto por Thissen *et al.* (1988) es uno de los que, en el marco de la TRI, se pueden aplicar con mayor facilidad dado que no requiere la igualación de parámetros de items entre grupo focal y el grupo de referencia (requisito imprescindible cuando se trabaja con las medidas exactas de área de Raju y con el estadístico de Lord) al estimarse conjuntamente los parámetros de ambos grupos.

Sin embargo, una limitación importante de las técnicas basadas en la TRI es precisamente la necesidad de elevados tamaños muestrales de los grupos focal y de referencia para asegurar una estimación apropiada de los parámetros del test y, por consiguiente, del DIF. La cuestión del tamaño muestral es relevante porque en la práctica los estudios aplicados de DIF disponen de pocos sujetos en el grupo focal, y en ocasiones tampoco el grupo de referencia es suficientemente numeroso. En este sentido, los procedimientos encuadrados en el apartado de métodos de invarianza condicional observada son más adecuados cuando se trabaja con grupos reducidos. De todos ellos, el estadístico MH resulta el más fácil de aplicar ya que no requiere algoritmos complejos de estimación y posiblemente es el más sencillo de comprender para profesionales con poco dominio de la estadística; este hecho, junto a que presenta un alto porcentaje de acuerdo, en la detección del DIF uniforme, con los procedimientos derivados de la *TRI* cuando se trabaja bajo el modelo de Rasch, lo ha llevado a ser el procedimiento seleccionado por el *Educational Testing Service* en Estados Unidos para la de-

tección de items con DIF en los programas de evaluación del rendimiento de los estudiantes norteamericanos. A pesar de ello, el procedimiento MH aplicado de forma estándar no resulta apropiado para detectar DIF no uniforme, presenta problemas de precisión cuando los items son muy difíciles y poco discriminativos y además no utiliza toda la información posible dado que la puntuación observada en el test se categoriza en k intervalos.

En caso de DIF no uniforme, el método logit, la RL y el procedimiento SIBTEST, constituyen las alternativas a tener en cuenta. El principal inconveniente de SIBTEST es que resulta complejo decidir bajo qué criterios se selecciona el subtest válido que no contenga items sesgados. Entre los modelos logit y la RL, es preferible esta última ya que trabaja con todos los niveles de habilidad sin pérdida de información; aunque RL es capaz de detectar DIF uniforme y no uniforme, cabe resaltar que en la identificación de DIF uniforme tiene menor potencia que el estadístico MH (Navas y Gómez, 1994; Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990).

Las aportaciones metodológicas que se han ido produciendo sobre diferentes técnicas de DIF y sobre la valoración de su eficacia relativa han tenido una influencia considerable en los sucesivos estudios empíricos llevados a cabo sobre el posible sesgo de distintos instrumentos de medida. En la Tabla 2 aparecen ordenadas, desde 1980 hasta la actualidad, las aplicaciones prácticas sobre el tema. Cabe resaltar, en primer lugar, que a partir del año 1985 ya no se utilizan procedimientos encuadrados en los denominados métodos pioneros que pueden considerarse obsoletos por cuanto, al no igualar a los grupos en el nivel de habilidad, no permiten distinguir entre efectos de impacto y DIF. Con respecto a las técnicas condicionales, la Tabla 2 pone de relieve que en las aplicaciones empíricas referenciadas están reflejadas prácticamente todas las técnicas expuestas en el presente trabajo, excepto las de muy reciente aparición; de todos modos, hay que destacar que los investigadores han seleccionado preferentemente aquellas basadas en la TRI (entre éstas, particularmente la χ^2 de Lord, las medidas del área y la G^2), el estadístico Mantel-Haenszel y el procedimiento estandarizado.

La mayor parte de los estudios empíricos se refieren a tests de conocimientos y aptitudes, pero se tratan también algunas escalas que miden constructos actitudinales y de personalidad y es de prever que el campo se vaya extendiendo en este sentido en el futuro. En estos trabajos se intenta determinar en qué grado diversas características del sujeto tienen un efecto no deseado en la medición. Como puede apreciarse en la Tabla 2, las variables de DIF más ampliamente analizadas hasta ahora son la raza y el sexo, por su indudable impacto político y social, particularmente en Estados Unidos que es donde se llevan a cabo la mayor parte de las investigaciones sobre DIF. Sin embargo, se constata también el estudio creciente de una gran disparidad de variables de tipo educativo, lingüístico, cultural, etc, que abren sin duda un amplio abanico de posibilidades poco explotadas por el momento; es de destacar en particular, por su prometedor futuro, la línea de investigación desarrollada por Hulin y posteriormente por Ellis en relación a la equivalencia de la medición de los tests traducidos con respecto al original y al grado en que los items con DIF entre las distintas versiones de un mismo test constituyen una fuente de información sobre posibles diferencias culturales.

TABLA 2. ESTUDIOS APLICADOS SOBRE DIF

Test	*Variable de* DIF	*Procedimiento*	*Autores*
Test de McCarthy, Lectura, Aritmética, Conocimiento y Discriminación de palabras,...	Raza	ANOVA	Reynolds (1980)
Matemáticas	Raza	Procedimiento de Linn y Harnisch	Linn y Harnisch (1981)
Test Iowa de Habilidades Básicas en Matemáticas (ITBS)	Sexo	ANOVA	Plake, Loyd y Hoover (1981)
Escala de Satisfacción laboral	Dominancia lingüística y Traducción de tests	Hulin *et al.* (1982)	Hulin, Drasgow y Komocar (1982)
Escalas de McCarthy	Raza	χ^2 de Pearson	Murray y Mishra (1983)
Escala de Inteligencia (WISC)	Raza	Correlaciones biseriales y de rango	Ross-Reynolds y Reschly (1983)
Peabody Picture Vocabulary Test-revised (PPVT-R)	Raza	ANOVA	Arguiewicz y Abel (1984)
Habilidades Generales (Matemáticas, Lenguaje, ...)	Raza y Sexo	Delta-plot, Correlación biserial, Estadístico de Scheuneman y Procedimiento de Linn y Harnisch	Hoover y Kolen (1984)
Test de aptitudes matemáticas	Nivel lector	Delta-plot, Procedimiento de Linn y Harnisch y χ^2 de Camilli	Ironson, Homan, Willis y Signer (1984)
Test de Aptitudes Escolares (SAT). Rendimientos de Matemáticas y Lengua	Sexo, Raza y Nivel educativo	Procedimiento Estandarizado	Dorans y Kulick (1986)
Escala de satisfacción laboral	Traducción de tests	Hulin *et al.* (1982)	Hulin y Mayer (1986)
Culpabilidad Sexual	Sexo	G^2 (TRI)	Thissen, Steinberg y Gerrard (1986)
SAT	Discapacidades físicas	Regresión logística	Bennett, Rock y Kaplan (1987)
Escala afectiva de autoconcepto	Raza	Análisis Factorial Confirmatorio	Benson (1987)
ACTM (*Assessment Mathematics Usage Test*)	Sexo	Procedimiento de Linn y Harnisch	Doolittle y Cleary (1987)

(continúa en la pág. 25)

(viene de la pág. 24)

Test	*Variable de DIF*	*Procedimiento*	*Autores*
Test de Aptitudes en Matemáticas y Lengua	Sexo y Raza	χ^2 de Lord	Drasgow (1987)
SAT	Raza	Procedimiento estandarizado	Schmitt (1988)
Matemáticas (sustracción de fracciones)	Estrategias de resolución de problemas	MH	Bennett, Rock y Novatoski (1989)
Test de Inteligencia	Cultura y Traducción de tests	χ^2 de Lord	Ellis (1989)
Cuestionario de Actitudes en Salud Mental	Cultura y Traducción de tests	χ^2 de Lord	Ellis, Minsel y Becker (1989)
Conocimiento y aptitudes generales	Raza	Medidas de área y MH en dos etapas	Hambleton y Rogers (1989)
Test de vocabulario	Raza	MH	Raju, Bode y Larsen (1989)
K-ABC (*Test of Intellectual Processing Skills*)	Raza y Sexo	Índice de Stricker	Willson, Nolan, Reynolds y Kamphans (1989)
National Assessment of Educational Progress (NAEP) (Historia)	Sexo, Raza y Bagaje educativo	MH	Zwick y Ercikan (1989)
SAT (Matemáticas)	Sexo	Modelo de Rasch	Becker (1990)
Test de Inteligencia	Raza	χ^2 de Lord	Ellis (1990)
Habilidades para la enseñanza	Raza	MH y Estadístico de Wright y Stone	Engelhard, Anderson y Gabrielson (1990)
SAT y GRE (Aptitud Verbal)	Raza	Procedimiento estandarizado	Freedle y Kostin (1990)
Matemáticas	Sexo	χ^2 de Camilli y correlaciones por rango	Kim, Plake, Wise y Novak (1990)
Comprensión lectora	Raza y Sexo	MH	Scheuneman y Gerritz (1990)
SAT	Raza	Procedimiento estandarizado	Schmitt y Dorans (1990)
Prueba de inglés	Idioma	Variación Delta-Plot	Sasaki (1991)
SAT	Sexo	G^2 (TRI)	Wainer, Sireci y Thissen (1991)
Matemáticas	Uso de calculadora	MH y χ^2 de Lord	Cohen y Kim (1992)

(continúa en la pág. 26)

(viene de la pág. 25)

Test	*Variable de DIF*	*Procedimiento*	*Autores*
Test de personalidad	Cultura y Traducción de tests	χ^2 de Lord	Ellis, Becker y Kimmel (1993)
Test de vocabulario	Sexo y Raza	Medidas de área, MH y χ^2 de Lord	Raju, Drasgow y Slinde (1993)
Test de ciencias	Sexo	MH iterativo y juicios de expertos	Sudweeks y Tolman (1993)
Escala de suicidio	Raza, Sexo y Salud Mental	Modelos loglineales	Dancer, Anderson y Derlin (1984)
Comprensión lectora y Matemáticas	Raza	Medidas de área de Kim y Cohen	Oshima, McGinty y Flowers (1984)
Matemáticas	Uso de calculadora	χ^2 de Lord, Medidas de Raju y G^2 (TRI)	Kim y Cohen (1995)
Comprensión lectora y razonamiento	Raza	G^2 (TRI)	Wainer (1995)

Conclusión

La exposición de las principales técnicas de DIF y de sus aplicaciones prácticas ha pretendido evidenciar la importancia actual del tema y ofrecer pautas para su correcta utilización.

Queda patente la relevancia del estudio del DIF de cara a evitar la posible falta de imparcialidad para ciertos sectores de la población que pueda concatenar el uso de tests con ítems que funcionan diferencialmente entre distintos grupos; tema que no sólo repercute a nivel educativo o laboral, sino que puede involucrar aspectos políticos y legales. Todo este afán por analizar y estudiar el DIF va encaminado hacia la obtención de «tests libres de cultura» que no discriminen en forma negativa a los miembros de unos determinados grupos sólo por el hecho de pertenecer a ellos. Es un tema crucial, por tanto, la identificación de aquellos ítems que puedan estar perjudicando a los individuos por pertenecer a un grupo u otro de la población a causa de su sexo, etnia, clase social, religión o contexto cultural; con su consecuente eliminación puede evitarse que miembros de grupos distintos en estas características pero de igual nivel en el atributo medido se vean beneficiados o perjudicados.

Los estudios sobre DIF se muestran como vehículos metodológicos de primera categoría para valorar la equidad en la evaluación con un determinado instrumento de medida, y la creciente cantidad de investigaciones sobre ítems o tests que funcionan diferencialmente pone en evidencia los esfuerzos por lograr

mediciones objetivas; de la importancia del tema no sólo dan cuenta las numerosas investigaciones metodológicas sobre métodos de detección sino también el análisis de sus consecuencias prácticas aplicadas a muy diversos contenidos. En general, y a la vista de la valoración de las técnicas de detección de DIF, en un estudio empírico resulta conveniente utilizar más de una técnica de detección de DIF antes de tomar una decisión acerca de si un ítem presenta o no DIF y, una vez que mediante algún procedimiento estadístico se ha probado que un determinado ítem funciona diferencialmente para dos o más grupos, es imprescindible estudiar las causas de ese DIF y encontrar una explicación teórica de la ocurrencia del mismo. Sin embargo, como puede apreciarse en la Tabla 2, pocos estudios aplicados utilizan una estrategia de convergencia entre técnicas y menos todavía abordan el estudio de las causas del funcionamiento diferencial encontrado; parece evidente que si bien el desarrollo metodológico del tema ha sido extenso y riguroso, en sus aplicaciones prácticas se está prácticamente en los comienzos y cabe pensar que se abre un futuro prolífico sobre el estudio del sesgo de los instrumentos de medida en el marco de la validez de constructo del test en cuestión que analice en profundidad las variables que originan la dificultad relativa de los ítems para distintos grupos.

No sólo es de gran interés detectar y eliminar del test aquellos ítems que muestran DIF con respecto a ciertos grupos de la población, sino conocer el porqué del funcionamiento diferencial de esos ítems. Así, una vez conocidos los factores causantes del DIF, los constructores de tests podrían prevenir con mayor facilidad la ocurrencia de ítems sesgados. Aunque el hallar esas causas es una tarea dura y problemática, algunos estudios han intentado esclarecer las posibles causas del funcionamiento diferencial de algunos ítems (Scheuneman, 1987; Skaggs y Lissitz, 1992). Como ya se ha expresado en la introducción, se entra con ello en el ámbito del sesgo entendido como fuente de invalidez o de error sistemático que se refleja en cómo un test mide a los miembros de un grupo particular (Camilli y Shepard, 1994).

Un análisis de DIF por sí solo no ofrece ninguna información del porqué los grupos se comportan de manera distinta con respecto a un ítem determinado, mientras que un estudio de sesgo pretende ofrecer una explicación de este comportamiento diferencial. En el estado actual de la investigación sobre el tema, dicha explicación recae básicamente en la posible multidimensionalidad de la medida (Ackerman, 1992; Oort, 1996; Shealy y Stout, 1993a, 1993b) y en la consiguiente identificación de qué otras variables mide el test, aparte de la variable principal; dichas variables serían de naturaleza espúrea y afectarían al nivel conseguido por los sujetos en la variable de interés.

Oort (1996) las denomina «violadoras de la unidimensionalidad» y plantea un modelo en el que es posible considerar simultáneamente varios factores que pueden influenciar la manera en que un sujeto responde a un ítem y que no sólo afectan al grupo de pertenencia, sino a la ocasión temporal en que la variable es medida, a la sensibilidad del ítem a una categoría de respuesta determinada, a la facilidad con que un rasgo encubre a otros, etc. El primer paso en la elaboración de un test válido es identificar a los posibles violadores de la unidimensionalidad, teniendo en cuenta la naturaleza del rasgo que se pretende medir,

el tipo de items con que se mide, los grupos a los que es aplicable, etc.; una vez operacionalizados estos potenciales factores de sesgo, puede comprobarse su influencia en las respuestas de los sujetos e interpretar desde la teoría por qué sujetos del mismo nivel de habilidad tienen diferentes probabilidades de responder correctamente a un ítem determinado.

En definitiva, cabe concluir que los estudios de DIF y, en la medida de lo posible, de sus causas han de convertirse en rutinarios tanto en la evaluación de tests existentes como en el desarrollo de nuevos instrumentos de medida. En este último aspecto, deben constituir una parte más del proceso de construcción de un test, como lo son el análisis de items o el estudio de la fiabilidad. Para evitar al máximo el sesgo conviene examinar cuidadosamente el contenido de los items por manos de expertos, detectar a posteriori los items que muestren DIF y analizar las posibles causas subyacentes; ello conllevará una óptima garantía de que realmente los items del test resultante miden por igual a todos los grupos de población que pueden ser evaluados con él.

Referencias

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

Angoff, W.H. (1982). Uses of difficulty and discrimination indices for detecting item bias. En R.A. Berk (Ed) *Handbook of methods for detecting item bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.

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

Argulewicz, E.N. y Abel, R.R. (1984). Internal evidence of bias in the PPVT-R for Anglo-American and Mexican-American children. *Journal of School Psychology, 22*, 299-303.

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

Becker, B.J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal, 27*, 65-87.

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.

Bennett, R.E. , Rock, D.A. y Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille Edition. *Journal of Educational Measurement, 26*, 67-79.

Benson, J. (1987). Detecting item bias in affective scales. *Educational and Psychological Measurement, 47*, 55-67.

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

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

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. y Shepard, L.A. (1987). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics, 12*, 87-99.

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 assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253-260.

Clauser, B., Mazor, K. 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.

Cleary, T.A. y Hilton, T.L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28*, 61-75.

Cohen, A.S. y Kim, S.H. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education, 5*, 303-320.

Cohen, A.S., Kim, S.H. y Subkoviak, M.J. (1991). Influence of prior distributions on detection of DIF. *Journal of Educational Measurement, 28*, 49-59.

Dancer, L.S., Anderson, A.J. y Derlin, R.L. (1994). Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. *Journal of Consulting and Clinical Psychology, 62*, 710-717.

Doolittle, A.E. y Cleary, T.A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*, 157-166.

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 assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N.J., Schmitt, A.P. y Bleinstein, C.A. (1992). The standarization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309-319.

Drasgow,F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.

Ellis, B.B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74*, 912-921.

Ellis, B.B. (1990). Assessing intelligence Cross-Nationally: A case for differential item functioning detection. *Intelligence, 14*, 61-78.

Ellis, B.B., Becker, P. y Kimmel, H.D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24*, 133-148.

Ellis, B.B., Minsel, B. y Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. Special Issue: Cross-cultural comparison of psychological data: Issues and pitfalls. *International Journal of Psychology, 4*, 665-684.

Engelhard, G.Jr., Anderson, D. y Gabrielson, S. (1990). An empirical comparison of Mantel-Haenszel and Rasch procedures for studying differential item functioning on teacher certification tests. *Journal of Research and Development in Education, 23*, 173-179.

Fidalgo, A.M. (1996). *Funcionamiento diferencial de los items:Procedimiento Mantel-Haenszel y modelos loglineales.* Tesis doctoral no publicada. Universidad de Oviedo.

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 items uniforme y no uniforme.* Comunicación presentada al IV Symposium de Metodología de las Ciencias del Comportamiento, La Manga, Murcia.

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

Freedle, R. y Kostin, I. (1990). Item difficulty of four verbal item types an index of differential item functioning for black and white examinees. *Journal of Educational Measurement, 27*, 329-343.

Green, D.R. y Draper, J.F. (1972). *Exploratory studies of bias in achievement tests.* Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Gómez, J. y Navas, M.J. (1996). Detección del funcionamiento diferencial de los items mediante regresión logística: Purificación paso a paso de la habilidad. *Psicológica, 17*, 397-411.

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. (1995). *Evaluación del funcionamiento diferencial del ítem en items dicotómicos y politómicos: un estudio comparativo.* Tesis doctoral no publicada. Universidad de Murcia.

Hidalgo, M.D. y López Pina (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 items. *Psicothema, 9*, 417-431.

Holland, P.W. y Rubin, D.B. (1982). *Test equating.* Princeton, NJ: Academic Press.

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.

Hoover, H.D. y Kolen, M.J. (1984). The reliability of six item bias indices. *Applied Psychological Measurement, 8*, 173-181.

Hulin, C.L., Drasgow, F. y Komocar, J. (1982). Application of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67*, 818-825.

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

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.

Ironson, G.H., Homan, S., Willis, R. y Signer, B. (1984). The validity of item bias techniques with math word problems. *Applied Psychological Measurement, 8*, 391-396.

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

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.

Kim, S.H. y Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51-66.

Kim, S.H. y 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*, 291-312.

Kim, S.H., Cohen, A.S. y Kim, H.O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*, 217-228.

Kim, H., Plake, B.S., Wise, S.L. y Novak, C.D. (1990). A longitudinal study of sex-related item bias in mathematics subtests of the California Achievement Test. *Applied Measurement in Education, 3*, 275-284.

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.

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.

Li, H. y Stout, W. (en prensa). A new procedure for detecting crossing DIF. *Psychometrika*.

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

López Pina, J.A., Hidalgo, M.D. y Sánchez, J. (1993). *Error tipo I de las pruebas χ^2 en el estudio del sesgo de los items*. Comunicación presentada al III Symposium de Metodología de las Ciencias del Comportamiento, Santiago de Compostela.

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.

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.

McLaughlin, M.E. y Drasgrow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161-173.

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.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300-307.

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.

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

Murray, A.M. y Mishra, S.P. (1983). Judgments of item bias in the McCarthy scales of children's abilities. *Hispanic Journal of Behavioral Sciences, 5*, 325-336.

Narayanan, P. y Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.

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

Oort, F.J. (1996). *Using restricted factor analysis in test construction*. Amsterdam: Universidad de Amsterdam.

Oshima, T.C., McGinty, D. y Flowers, C.P. (1994). Differential item functioning for a test with a cuttoff-score: Use of limited closed-interval measures. *Applied Measurement in Education, 7*, 195-209.

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.

Plake, B.S. y Hoover, H.D. (1979). An analytical method of identifying biased test items. *The Journal of Experimental Education, 48*, 153-154.

Plake, B.S., Loyd, B.H. y Hoover, H.D. (1981). Sex differences in mathematics components of the Iowa Tests of Basic Skills. *Psychology of Women Quartely, 5*, 780-784.

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 assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2*, 1-13.

Raju, N.S., Drasgow, F. y Slinde, J.A. (1993). An empirical comparision of the area methods, Lord's chi-square tests, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53*, 301-314.

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

Reynolds, C.R. (1980). An examination for bias in a preschool test battery across race and sex. *Journal of Educational Measurement, 17*, 137-146.

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.

Ross-Reynolds, J. y Reschly, D.J. (1983). An investigation of item bias on the WISC-R with four sociocultural groups. *Journal of Consulting and Clinical Psychology, 51*, 144-146.

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

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.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*, 95-111.

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

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

Scheuneman, J.D. y Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*, 109-131.

Schmitt, A.P. (1988). Language and cultural characteristics that explain differential item functioning for hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement, 25*, 1-13.

Schmitt, A.P. y Dorans, N.J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*, 67-81.

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

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

Skaggs, G. y Lissitz, R.W. (1992). The consistency of detecting item bias across of different test administration: implications of another failure. *Journal of Educational Measurement, 29*, 227-242.

Sudweeks, R.R. y Tolman, R.R. (1993). Empirical vs subjetive procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching, 30*, 3-19.

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

Tatsuoka, K.K., Linn, R.L. Tatsuoka, M.M. y Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25*, 301-319.

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.

Uttaro, T. y Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

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

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Tests as an example. *Applied Measurement in Education, 8*, 157-186.

Wainer, H., Sireci, S.G. y Thissen, D. (1991). Differential item functioning: definitions and detection. *Journal of Educational Measurement, 28*, 197-219.

Westers, P. y Kelderman, H. (1991). Examining differential item functioning due to item difficulty and alternative attractiveness. *Psychometrika, 57*, 107-118.

Willson, V.L., Nolan, R.F., Reynolds, C.R. y Kamphan, R.W. (1989). Race and gender effects on item functioning on the Kaufman Assesment Battery for Children. *Journal of School Psychology, 27*, 289-296.

Zwick, R. y Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.