ACTA COLOMBIANA DE PSICOLOGÍA 5, 41-61-01

FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS: UNA REVISIÓN CONCEPTUAL Y METODOLÓGICA

AURA NIDIA HERRERA ROJAS*
RICARDO SÁNCHEZ PEDRAZA**
UNIVERSIDAD NACIONAL DE COLOMBIA

JUANA GÓMEZ BENITO***
UNIVERSIDAD DE BARCELONA

Today, identifying biased item tests is an inevitable task in the process of construction and validation of psychological tests. Thus, it is important to understand the concepts related to Differential Items Functioning (DIF) and the methods for its detection. This paper reviews the basic concepts related to DIF within the context of tests construct validity. In addition, some important methods, commonly used for DIF detection, are described emphasizing the advantages and limitations for its practical use.
Key word: DIF, Bias Item, Bias Test.

Uno de los argumentos frecuentes de quienes participaron en lo que Anastasi (1974, p. 565) denominó la "*revuelta anti-tests*", era el hecho de que las pruebas psicológicas discriminaban grupos minoritarios puesto que sus resultados, y por ende, las decisiones que se tomaban con base en los mismos, resultaban injustas para algunos de estos grupos. Esa percepción de injusticia se basaba en la observación de una diferencia sistemática en los puntajes de las pruebas, entre personas pertenecientes a distintos grupos según género, raza, clase socioeconómica o cultura; resultado que desfavorecía a alguno(s) de los grupos, generalmente minoritarios, en comparación con otros. La existencia de esas diferencias no era sin embargo, desconocida por los psicólogos y educadores de la época; desde comienzos de siglo William Stern (1914, citado por Camilli y Shepard, 1994) había mostrado que el rendimiento en pruebas de inteligencia variaba según la clase social, Binet y Simon (1916, citado por Camilli y Shepard, 1994) habían eliminado algunos ítems en la nueva versión de su prueba de inteligencia porque eran sensibles a *efectos culturales* y a mediados de siglo Eelles, Havighurst, Herrick y Tyler (1951) habían mostrado cómo algunos ítems presentaban lo que para el momento se denominó *sesgo cultural* (Camilli y Shepard, 1994; Fidalgo, 1996; Ferreres, 1998). Este último trabajo junto con el de Jensen (1969) merecen un

* E-mail: anherrer@bacata.usc.unal.edu.co
** E-mail: *risanche@bacata.usc.unal.edu.co*
*** E-mail: *jgomez@psi.ub.es*

lugar destacado a la hora de realizar una revisión histórica y conceptual de lo que hoy se conoce como funcionamiento diferencial de los ítems (DIF, abreviatura de *Differential Item Functioning*).

El primero de estos dos trabajos es considerado por la mayoría de autores sobre el tema, como la investigación pionera sobre sesgo de los ítems de pruebas psicológicas; más allá de la discusión sobre el sesgo de las pruebas en su totalidad, Eelles, Havighurst, Herrick y Tyler (1951) mostraron empíricamente y con un gran número de ítems de pruebas de inteligencia, que algunos de ellos se dejaban afectar por diferencias culturales. Aunque sus hallazgos habían aparecido publicados una década antes, la discusión sobre el sesgo de las pruebas alcanzó su punto más alto durante los sesenta; este hecho se entiende si se tiene en cuenta la dimensión que tomó el movimiento por la defensa de los derechos civiles y por consiguiente, de la igualdad de oportunidades educativas y laborales; contextos en los que las pruebas psicológicas eran ampliamente usadas y sus resultados constituían uno de los argumentos más fuertes para la toma de decisiones relacionadas con asignación de cupos (Cole, 1993; Fidalgo, 1996).

En esta discusión, sin embargo, la noción de *sesgo* se asociaba a cualquier diferencia sistemática entre grupos diferentes, en cuanto a los resultados de las pruebas y en consecuencia, el término *sesgo* tenía una connotación negativa equiparable con *injusticia*, *parcialidad* e *inequidad* contra los grupos minoritarios o menos favorecidos. Angoff (1993), Cole (1993) y Holland y Wainer (1993) parecen estar de acuerdo en que esta asociación, dominante durante la década de los sesenta y que tuvo amplias implicaciones de tipo social, se debió en gran parte a un conflicto semántico y a una confusión en el uso del lenguaje común y del lenguaje técnico: público y psicólogos estaban usando el mismo término –*sesgo*– pero para los primeros estaba cargado de contenido social y político con una connotación claramente negativa, mientras que para los segundos estaba cargado de contenido técnico, y aunque la connotación no era buena, hacía referencia básicamente a '*características técnicas no óptimas*' (Cole, 1993, p.27) y no a injusticia social.

Fidalgo (1996) por su parte, identifica en esta discusión la *falacia igualitaria* (p.376) basada en el supuesto de la igualdad entre los hombres, de manera que las pruebas o cualquier instrumento que pusiera en evidencia diferencias entre grupos humanos, resultaba discriminatorio y sesgado. Sorprendentemente, un claro ejemplo de esta confusión se presentó en Estados Unidos, casi dos décadas después –en 1984– cuando se había avanzado en la claridad conceptual y en el desarrollo de técnicas de detección del DIF. *Golden Rule Insurance Company* demandó al Departamento de Seguros de Illinois y al ETS (*Educational Testing Service*) con el argumento de que algunas pruebas desarrolladas por el ETS para el Estado de Illinois estaban sesgadas contra los negros. El famoso caso derivó en lo que se conoció como la regla de oro (*Golden rule*) consistente en que el ETS no incluiría en sus pruebas, ítems que mostraran diferencias de dificultad[1] superiores a *0,15* entre blancos y negros. Aunque la

[1] Dificultad evaluada como la proporción observada de aciertos en la pregunta, según se define en la teoría clásica de los tests.

decisión se tomó en una negociación fuera de la Corte y en ese sentido no constituyó antecedente legal, académicos (Bond, 1987; Faggen, 1987; Linn y Drasgow, 1987, entre otros) y la misma Asociación Americana de Psicología (American Psychological Association - APA) debieron pronunciarse al respecto debido a las posibles implicaciones de tipo legal y político (Anastasi y Urbina, 1998) y a algunos intentos por generalizar la regla en otros contextos. Lim y Drasgow (1990) citan algunos de estos intentos.

En medio de la acalorada discusión de finales de los 60's aparece publicado el segundo trabajo citado antes como importante en el desarrollo de lo que hoy se conoce como DIF, por la enorme polémica que desató y su efecto sobre el desarrollo de técnicas para su detección -el de Jensen (1969). El autor argumentaba que la inteligencia era heredada y que en consecuencia, las diferencias observadas en las pruebas entre grupos raciales podían explicarse genéticamente. Estos argumentos avivaron la discusión entre los partidarios de la explicación genética y los partidarios de los determinantes ambientales y sociales de las diferencias en C.I. Los últimos atribuían en gran medida las diferencias entre grupos, al sesgo de las pruebas. En lo que tiene que ver con psicometría propiamente, sin embargo, la importancia del trabajo estuvo en que puso de manifiesto la necesidad de evaluar hasta qué punto las diferencias observadas en las pruebas se debían a las características reales de los grupos o a artificios generados por el instrumento mismo, lo cual implicaba además de hacer claridad conceptual, generar técnicas que permitieran evaluar el posible efecto de las pruebas mismas y de los ítems que la componen, en las diferencias entre grupos. Muñiz (1998) hace notar cómo las publicaciones psicométricas especializadas de los 50's y los 60's y la edición de 1966 de los *Standards for Educational and Psychological Test and Manuals* ignoran por completo el tema, y es sólo a partir de los 70 cuando la comunidad psicométrica se apropia de la discusión que hasta el momento se había mantenido en las esferas legal, política, social y de la teoría psicológica.

En las últimas tres décadas y tras la aparición del polémico artículo de Jensen (1969), se ha llegado a conceptualizaciones mucho más precisas y se han generado múltiples estrategias de estimación del DIF. Sin lugar a dudas, una propuesta que contribuyó a hacer claridad conceptual y superar el conflicto semántico mencionado antes fue la de Holland y Thayer (1988) quienes sugirieron cambiar el término *sesgo* por el de *Funcionamiento diferencial de los ítems* (DIF). Así, el primer término se usaría para referirse a un '*juicio informado*' (Holland y Wainer, 1993, p. xiv) que además de la información estadística sobre el ítem, tomara en cuenta el objetivo de la prueba y la información de tipo histórico, social o cultural que posiblemente explicara su funcionamiento. Aunque la nueva expresión resultaba en opinión de Angoff (1993) más larga y menos comprensible, rápidamente se fue generalizando a través del uso de su abreviatura DIF, en las publicaciones especializadas, para referirse a la observación desapasionada del hecho de que algunos ítems pueden mostrar propiedades psicométricas diferentes para grupos diferentes.

Además del cambio de terminología, en el mismo período de tiempo se han publicado gran cantidad de trabajos sobre el tema, se han identificado categorías diferentes de DIF y se han propuesto múltiples alter-

nativas metodológicas de estimación, incluyendo el juicio humano (Anastasi y Urbina, 1998; Ferreres, 1998; Muñiz, 1998; Muñiz, 1997). La literatura especializada de las últimas décadas muestra un número considerable de esfuerzos por identificar las ventajas y limitaciones de las diferentes técnicas de estimación, así como por encontrar las condiciones en las cuales resultan más o menos adecuados. Este artículo se ocupa de presentar una revisión del concepto de Funcionamiento Diferencial de los Ítems y de las técnicas más frecuentemente usadas en la actualidad para su detección. Esta última parte, sin embargo, se limita a una somera descripción de los métodos más usados para la estimación del DIF en ítems de tipo dicotómico.

FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

Uno de los ejemplos más ilustrativos para entender la noción de sesgo o funcionamiento diferencial, es el que presenta Muñiz (1998): "*un metro estará sistemáticamente sesgado si no proporciona la misma medida para dos objetos o clases de objetos que de hecho miden lo mismo, sino que sistemáticamente perjudica a uno de ellos*" (p.236). Esta afirmación tiene varias implicaciones que permiten aclarar la noción de funcionamiento diferencial. En primer lugar, se habla de funcionamiento diferencial cuando se tienen por lo menos dos grupos de comparación que se denominan generalmente grupo de referencia y grupo focal. Por otra parte, para afirmar que un instrumento o un ítem funcionan diferencialmente es necesario tener alguna evidencia de que los grupos que se comparan tienen el mismo nivel del atributo medido y finalmente, lo que define el funcionamiento diferencial es el hecho de que los resultados que arroja el instrumento son sistemáticamente diferentes entre los grupos. Sin embargo, el *metro* puede hacer referencia a una prueba como un todo o a los elementos que la componen –ítems-.

La presentación del aparte anterior muestra que la discusión original acerca del *sesgo*, hacía referencia básicamente a las diferencias observadas en los resultados de las pruebas como un todo más que en lo que hoy se conoce como funcionamiento diferencial de los ítems en particular. Muchos autores han definido sesgo en las pruebas como una clase de *invalidez*, sin embargo, la validez involucra tanto características internas o propias de la prueba –validez de constructo- como aspectos externos al instrumento relacionados con el uso del mismo o el contexto de aplicación –validez predictiva-. Para Shepard (1982) hay un continuo de validez, en un extremo se encuentran las pruebas que miden lo que pretenden y lo hacen con la misma precisión para todos los grupos y en el otro extremo están las pruebas cuya validez implica tanto aspectos de tipo social y ético como argumentos de tipo científico y técnico, pasando por aquellas que muestran igual validez predictiva en contextos particulares.

El funcionamiento diferencial de las pruebas se ha clasificado generalmente como *sesgo* en la validez de constructo y *sesgo* en la validez predictiva. Este último implica necesariamente examinar la relación entre los puntajes de la prueba y una medida de criterio externo, de manera que se considera que una prueba tiene sesgo en cuanto a su validez predictiva si la predicción que se obtiene con base en sus puntajes no se

hace con el "menor error aleatorio posible o si hay un error constante" (Reynolds, 1982 p. 216) en la misma, en función del grupo. Su detección entonces, centra la atención en el análisis tanto de la pendiente como del intercepto de las ecuaciones de regresión obtenidas para los grupos de interés. Tanto en Reynolds (1982) como en Anastasi y Urbina (1998) se puede encontrar una exposición más detallada de estos procedimientos. De otra parte, se considera que una prueba tiene funcionamiento diferencial en cuanto a su validez de constructo, cuando mide algo diferente para diferentes grupos o, si mide lo mismo, lo hace con diferentes niveles de precisión (Reynolds, 1982). Este tipo de *invalidez* de las pruebas implica, entre otras cosas, el análisis de los ítems que componen la prueba y la evaluación de su coherencia lógica y teórica con el constructo que pretende medir; en consecuencia, está estrechamente relacionado con el funcionamiento diferencial de los ítems, tema del presente trabajo. En lo que resta, nos ocuparemos de éste y los métodos de detección.

Si al comparar dos grupos diferentes en cuanto a género, etnia, cultura o nivel socioeconómico, etc. se observan diferencias en la probabilidad de acertar en un ítem de una prueba, la primera pregunta relevante es ¿tales resultados se deben a diferencias reales en las magnitudes del atributo que pretende medir la prueba, o a otro tipo de variables relacionadas con la ejecución en el ítem pero diferentes al objeto de medida?. La identificación de la posible fuente de variación entre grupos permite distinguir el *Funcionamiento Diferencial* (DIF) del *Impacto* de los ítems, también conocido como *Impacto adverso* (Camilli y Shepard, 1994) o *diferencias válidas* (Van de Vijver y Leung, 1997). Cuando las diferencias observadas entre los grupos se deben a diferencias en la magnitud de atributo medido, se habla de impacto y cuando se deben a otras variables asociadas a los grupos comparados, se habla de DIF (Gómez e Hidalgo, 1997). Esta es una primera precisión conceptual importante por cuanto tiene amplias implicaciones de diferente orden; el caso antes mencionado que dio origen a la regla de oro (*Golden rule*) es un claro ejemplo de confusión entre DIF e impacto. Para Angoff (1993) allí hay un claro sesgo pero en el tratamiento que históricamente habían tenido los negros en la sociedad norteamericana, el cual producía condiciones educativas diferentes que conducían a desempeño más pobre en comparación con los blancos; las pruebas entonces recogen esa información que se manifiesta en diferencias de puntajes. Varios autores (Anastasi y Urbina, 1998; Angoff, 1993, entre otros) advierten que tal confusión puede conducir a que algunos ítems con alto poder de discriminación entre grupos de diferente nivel de ejecución, sean rechazados por ser sesgados, en detrimento de la confiabilidad y la validez del instrumento.

Una segunda distinción de orden conceptual es la que existe entre *DIF* y *sesgo*. Como ha podido verse, en los comienzos de la discusión el término que se utilizó para referirse al funcionamiento diferencial fue el segundo, y las diferentes connotaciones del mismo fueron en buena medida, causa de la gran confusión en torno al tema. Hoy, después de la propuesta de Holland y Thayer (1988), parece haber acuerdo en la necesidad de diferenciarlos claramente: El DIF hace referencia al hecho objetivo de que la probabilidad de acertar en un ítem cambia en función del grupo de pertenencia, es de-

cir, se centra en los procedimientos de tipo matemático y estadístico para identificar aquellos ítems que tienen diferente funcionamiento entre personas con igual magnitud del atributo medido pero que pertenecen a grupos diferentes. El sesgo incluye además, la identificación de las razones para que eso ocurra, lo cual trasciende los alcances de dichos métodos (Gómez e Hidalgo, 1997; Camilli y Shepard, 1994); el análisis de sesgo, que constituye la meta en la mayoría de trabajos aplicados, implica entonces, además de la identificación de los ítems que presentan DIF, la identificación del factor o factores que lo producen y la discusión teórica acerca de la relevancia de tales factores en el constructo que pretende medir la prueba.

El funcionamiento diferencial de un ítem puede apreciarse mediante la diferencia de probabilidad de éxito en el mismo, entre personas con igual magnitud del atributo medido, pero pertenecientes a diferentes grupos. En la figura 1 se representan las curvas características de cuatro ítems (CCI), esto es, la probabilidad de acertar al ítem (eje y) en función de la magnitud de atributo medido por la prueba (eje x). Las curvas del ítem 1 (figura 1a) son muy similares para los dos grupos de manera que no existe una diferencia en la probabilidad de acierto en el ítem entre los dos grupos comparados, sino que ésta cambia en función de la magnitud de atributo medido, se trata de un ítem que no presenta DIF. En las tres curvas restantes se puede apreciar diferencias en la probabilidad de acierto dada una magnitud de atributo, pero el comportamiento de las curvas es diferente entre ellas, lo que representa diferentes tipos de DIF.

Mellenbergh (1982) distingue el DIF uniforme del no uniforme, dependiendo de si existe interacción entre la magnitud de atributo que se pretende medir y la variable que identifica los grupos. Cuando no existe dicha interacción, es decir, cuando la magnitud de diferencia entre los grupos es similar en todos los niveles del atributo medido, el ítem presenta DIF uniforme; por el contrario, si existe interacción la diferencia en probabilidad de acertar al ítem cambia para diferentes niveles de atributo y existe DIF no uniforme (Gómez e Hidalgo, 1997; Ferreres, 1998; Prieto Marañón, Barbero García y San Luis Costas, 1997). En la figura 1b la probabilidad de acertar al ítem es mayor para el grupo 2 que para el grupo 1, de manera similar en todos los niveles de atributo: este ítem presenta DIF uniforme. Por el contrario, las curvas de las figuras 1c y 1d se cruzan en algún punto del eje x, lo cual muestra que la probabilidad de acierto es mayor para un grupo en niveles bajos de magnitud de atributo pero tal diferencia se invierte para niveles altos del mismo: estos dos ítems presentan DIF no uniforme.

Pero las curvas de las figuras 1c y 1d también muestran diferencias entre sí. En la primera, las curvas se cruzan aproximadamente a la altura de *0,5* en eje *y,* y la diferencia básica entre las dos es que una de ellas (la del grupo 1) tiene mayor pendiente que la otra; este ítem presenta dificultad similar para los dos grupos pero discriminación diferente, siendo más discriminativo para el grupo 1; se trata de un ítem que presenta DIF no uniforme propiamente dicho. Finalmente, en la figura 1d, las curvas se cruzan en un nivel muy bajo de magnitud de atributo y además, tienen pendiente diferente. Esto hace que la magnitud de atributo necesaria para tener una probabilidad de éxito igual a *0,5* sea mayor para el grupo 2 que para el grupo 1 y, además, la discri-

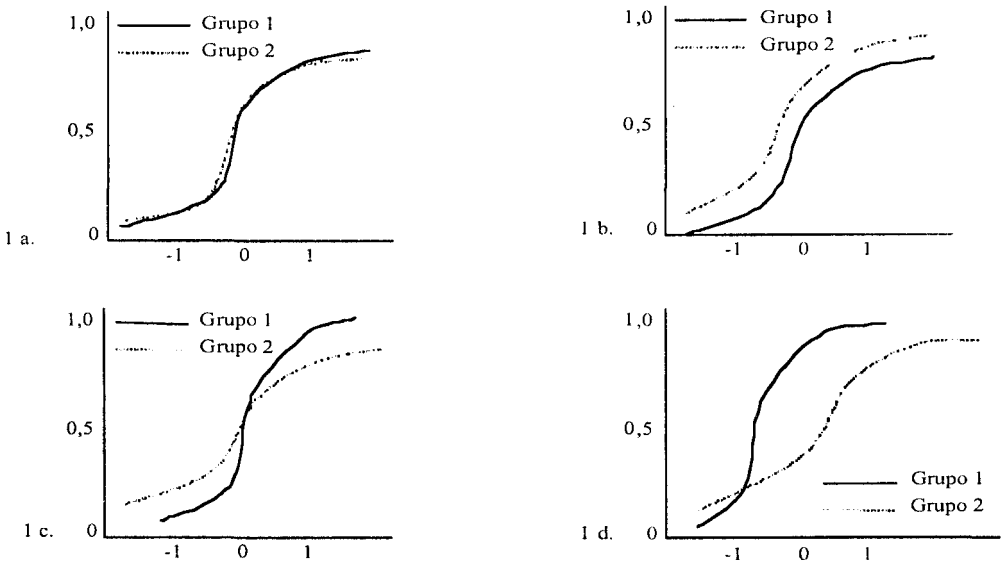


Figura 1: curva característica de 4 ítemes diferentes:(1a) ítem sin DIF (1b) ítem con DIF uniforme (1c) items con DIF no uniforme (1d) ítem con DIF mixto.

minación del ítem también es diferente para los dos grupos. Puede verse que aunque la diferencia en probabilidad de acierto también se invierte, los efectos no se anulan. Este ítem presenta DIF mixto.

En síntesis, en ambos ítems (figuras 1c y 1d) hay interacción entre la magnitud del atributo y el grupo pero en el primero el ítem resulta igualmente difícil y de diferente discriminación para los dos grupos, mientras que el segundo ítem tiene dificultad y discriminación diferente para los dos grupos. La interacción entre grupo y magnitud de atributo puede tener dos efectos observables en los parámetros de la Curva Característica del Ítem (CCI): el primero, conocido como DIF no uniforme, se da cuando el parámetro de dificultad no varía para los grupos pero el parámetro de discriminación si lo hace; lo cual produce que los efectos se anulen. El segundo, llamado DIF mixto, ocurre cuando ambos parámetros (dificultad y discriminación) cambian entre los grupos. Sobre estos parámetros se volverá más adelante al revisar los procedimientos de detección de DIF basados en la Teoría de Respuesta al Ítem (TRI).

MÉTODOS DE ESTIMACIÓN DE DIF

La revisión de la literatura sobre el tema muestra una diversidad de propuestas metodológicas para la estimación del DIF y de clasificaciones de las mismas. Camilli, y Shepard (1994) los clasifican en tres categorías: a) los métodos basados en el análisis de varianza y la Teoría Clásica de los Test (TCT), b) los que se basan en TR y c) los que se basan en el análisis de tablas de contingencia. Siguiendo a Millsap y Everson (1993), Gómez e Hidalgo (1997) clasifican

los métodos para detección de DIF de ítems dicótomos en métodos no condicionales y condicionales distinguiendo, dentro de la segunda categoría, los métodos de invarianza condicional observada y no observada. De otra parte, Fidalgo (1997) presenta dos grandes categorías: los procedimientos que no especifican un modelo de medida y los que se basan en la TRI. En Ferreres (1998) se puede encontrar una revisión de algunas clasificaciones que obedecen a criterios diferentes. En adelante nos ocuparemos de presentar una breve descripción de las principales técnicas que se han aplicado o que están en uso para evaluación de DIF, teniendo como guía la casificación de Camilli y Shepard (1994), como se ilustra en la Tabla 1.

TÉCNICAS BASADAS EN TCT Y ANOVA

Un primer grupo de procedimientos incluye las técnicas basadas en la TCT y en ANOVA, que también se ubicarían en la primera categoría de Gómez e Hidalgo (1997) ya que no efectúan ajustes en los grupos comparados en relación con la magnitud de atributo medido. Estos métodos, que constituyeron las primeras propuestas metodológicas, presentan la gran limitación de que pueden confundir el DIF con el impacto, precisamente porque no controlan la magnitud de atributo. Así, una diferencia significativa entre los grupos analizados en cuanto a la probabilidad de acertar en un ítem, puede conducir al rechazo de algunos elementos que tienen alta capacidad de discriminación entre personas con alto y bajo nivel de atributo. Esta razón ha hecho que estas técnicas de análisis hayan caído rápidamente en desuso (Gómez e Hidalgo, 1997). Solo se mencionarán aquí brevemente, cuatro de ellas: el índice transformado de dificultad, éste índice ajustado, el análisis de varianza y los métodos de correlación.

El índice transformado de dificultad (ITDI)

Se basa en el supuesto de que el DIF se manifiesta a través de una dificultad diferencial entre dos grupos y en conse-

Tabla 1. Una clasificación de técnicas para la detección del DIF

Categoría	Técnicas
Técnicas basadas en Teoría Clásica de los Test (TCT) y Análisis de varianza (ANOVA)	Índice transformado de dificultad (ITDI) ITDI ajustado Análisis de varianza Técnicas basadas en correlaciones
Técnicas basabas en el análisis de tablas de contingencias	X^2 de los aciertos Mantel-Haenszel Método de estandarización Modelos log-lineales Regresión logística
Técnicas basadas en la Teoría de Respuesta al Ítem (TRI)	Comparación de áreas Comparación de parámetros: X^2 de Lord Comparación de modelos

cuencia, simplemente ubicando los ítems con las mayores diferencias en la proporción de aciertos, se detectarán los ítems con DIF. Teniendo en cuenta este supuesto, se transforman las proporciones de acierto, *p*, en puntuaciones *Z* normalizadas correspondientes al *(1-p)ésimo* percentil. Posteriormente se efectúa una segunda transformación para obtener una distribución con media *0* y desviación estándar *1*. Los valores transformados para cada uno de los grupos se ubican en un dispersograma, de tal modo que si todos los ítem tienen la misma dificultad, los puntos caerán en una recta. En realidad los puntos caen en una nube de la cual puede calcularse un eje principal. De esta manera puede calcularse un valor de IDTI que es la distancia perpendicular desde el punto de un ítem hasta el eje principal. Valores altos de IDTI señalarían la presencia de DIF (Angoff, 1982). Como se comentó al inicio de este apartado, esta técnica confunde impacto y DIF, de manera que los ítems con buena capacidad de discriminación son señalados de tener DIF.

El ITDI ajustado

Busca superar las anteriores limitaciones. Si los grupos comparados fueran iguales en el desempeño total de la prueba el IDTI sería una técnica eficaz; sin embargo, para alcanzar este supuesto, no solo se requeriría que se igualaran las medias en el puntaje, sino que las distribuciones en el puntaje de cada uno de los grupos fueran semejantes. Angoff (1982) propuso una estrategia para manejar las limitaciones del método ITDI que consiste en emparejar los grupos con base en un criterio externo. Sin embargo, además de no resolver por completo las limitaciones del método, en la práctica dicho criterio externo no se tiene disponible. Otro ajuste también propuesto por Angoff se basa en las correlaciones ítem–prueba. Shepard (1984) propone el ajuste mediante el cálculo de índices ITDI residualizados. En la práctica ninguno de dichos métodos ha demostrado detectar mejor DIF que los métodos basados en tablas de contingencia que se presentarán más adelante.

El Análisis de varianza

Fue la técnica de elección para detección de DIF hasta comienzos de los años 80. Consiste en un análisis de varianza (ANOVA) de medidas repetidas de dos factores representados por el grupo y por el ítem. Este método se afecta por la dificultad media del ítem y por el impacto entre grupos. Debido a los altos valores de error tipo I y tipo II que genera, es una técnica que se ha abandonado (Camilli y Shepard, 1994, Gómez e Hidalgo, 1997).

De otra parte, se han propuesto varias *técnicas basadas en correlaciones*. Una de éstas consiste en ordenar los ítems de acuerdo con el nivel de dificultad, dentro de cada grupo y posteriormente correlacionar los rangos. Valores de correlación cercanos a *1* suponen que los ítems están midiendo un mismo atributo en ambos grupos; en caso contrario debe considerarse DIF. Otro método de correlación busca detectar cómo se comportan los índices de discriminación de los ítems en grupos diferentes, aplicando la correlación biserial-puntual; Sin embargo, ésta no es una técnica recomendada para la evaluación de DIF ya que la dificultad del ítem afecta este tipo de correlación.

MÉTODOS BASADOS EN TABLAS DE CONTINGENCIA

Dentro de los métodos basados en el análisis de tablas de contingencia se pueden distinguir dos enfoques: los que se fundamentan en la prueba hipótesis sobre la igualdad de proporciones y los que generan modelos para el análisis de variables categóricas. Dentro de los que prueban hipótesis sobre igualdad de proporciones se encuentran el X^2 de los aciertos, el Mantel-Haenszel y el método de estandarización; mientras que dentro de los últimos se pueden incluir los modelos loglineales y la regresión logística. Los primeros se fundamentan en que el DIF se detecta explorando una hipótesis nula relacionada con igualdad de proporciones de aciertos entre los grupos, controlando la magnitud de atributo medido. Si este último se fracciona de manera que genere diferentes estratos, para cada uno de estos estratos se tendrá una tabla de contingencias con la estructura que se ilustrada en la tabla 2.

Así, a_i representa el número de examinados del grupo de referencia que acertaron en el ítem y tienen la magnitud de atributo *i*, etc. La información completa sobre los aciertos y fallos para los dos grupos en cada ítem se tiene un tantas tablas bidimensionales como estratos según la magnitud de atributo medido. A partir de esta estructura básica se han desarrollado las diferentes técnicas antes mencionadas.

El X^2 *de los aciertos* parte del supuesto de que si, dentro de cada estrato, la proporción de aciertos es igual en cada uno de los grupos, se puede descartar la presencia de DIF. En este sentido Scheuneman (1979) propuso probar la hipótesis de igualdad de proporciones de los aciertos, mediante un estadístico X^2 con $(k-1)(r-1)$ con grados de libertad, siendo k el número de estratos y r el número de grupos. Para Baker (1981) este método, al considerar solamente los aciertos, puede dejarse afectar por distorsiones, sobre todo si existe impacto entre los grupos; caso en el cual los resultados pueden depender de qué tan similares sean los tamaños de los dos grupos. En estas condiciones, no resulta claro que la distribución que siga sea realmente la X^2.

Ironson (1982) cita un reporte de Camilli, quien hace un ajuste a la propuesta de Scheuneman calculando el valor X^2 tanto de aciertos como de fallos y evaluando la suma de éstos como una X^2 con $K(r-1)$ grados de libertad. Ambos métodos tienen el inconveniente de que no pueden detectar

Tabla 2. Estructura de una tabla de contingencia correspondiente a un nivel de magnitud de atributo *i*.

Grupo	Aciertos en el ítem		Total
	1	*0*	
Referencia	a_i	b_i	$N1 = a_i+b_i$
Focal	c_i	d_i	$N2 = c_i+d_i$
Total	$M1 = a_i+c_i$	$M2 = b_i+d_i$	n_i

DIF no uniforme y además se inestabilizan con valores bajos dentro de las celdas o ante la presencia de impacto en los grupos.

El *estadístico de Mantel–Haenszel* (1959) permite manejar de manera más eficiente los distintos niveles de habilidad como una variable de control. Este método permite evaluar y describir cómo la relación entre variables se modifica por la presencia de una variable externa, que es una variable categórica con *k* estratos. El estadístico de Mantel –Haenszel se puede presentar en dos componentes: el X^2 de Mantel-Haenszel y el *Odds Ratio* (OR= combinado de Mantel-Haenszel. El primero combina la información de las tablas correspondientes a cada uno de los estratos, generando un estadístico que mide la asociación global entre las dos variables –grupo R o F y puntaje dicotómico en el ítem- dado que los OR de los diferentes estratos son homogéneos, es decir que no hay interacción (Selvin, 1996). Para su cálculo, el procedimiento es el siguiente (Rosner, 1995): a) sumar los valores observados en la celda (1, 1) de cada uno de los estratos: $O = \sum_{i=1}^{k} O_i$, b) sumar los valores esperados en la celda (1, 1) de cada uno de los estratos, utilizando la función de valor esperado de una distribución hipergeométrica: $E = \sum_{i=1}^{k} E_i$, c) hacer la sumatoria de las varianzas V_i, asumiendo una distribución hipergeométrica: $V = \sum_{i=1}^{k} V_i$ y d) calcular el valor de prueba que es una diferencia de valores observados y esperados, haciendo corrección por continuidad: $\chi^2_{MH} = \frac{(|O-E|-0.5)^2}{V}$. Este estadístico tiene una distribución X^2 con *1* grado de libertad bajo la hipótesis nula de no asociación.

Una vez se ha establecido la significación de la asociación, puede estimarse la fuerza de la asociación. Si existe un OR común, éste puede estimarse (Kahn y Sempos, 1989). Esto se lleva a cabo mediante el cálculo del OR de Mantel-Haenszel: OR_{MH}. Esta medida da la fuerza de la asociación entre *2* variables, habiendo controlado el efecto de la variable de confusión:

$$OR_{MH} = \frac{\sum_{i=1}^{k} a_i d_i / n_i}{\sum_{i=1}^{k} b_i c_i / n_i}$$

Si el OR no es significativamente diferente de 1 puede concluirse que el ítem no tiene DIF. En este caso, responder acertadamente el ítem es igual de frecuente en el grupo R y en el grupo F, habiendo controlado por el efecto de otra variable que se relaciona con el test y con el grupo (nivel de habilidad). Si el criterio que se utiliza para seleccionar la variable de control y así establecer los niveles es la puntuación total de la prueba, debe hacerse un ajuste al estadístico para evitar problemas de circularidad, ya que los ítems con DIF también están contribuyendo a ese puntaje total. Los métodos propuestos incluyen practicar etapas sucesivas de purificación de ítems.

De acuerdo con el *método de estandarización*, un ítem presenta DIF cuando el desempeño esperado en él (*E*) difiere para individuos de grupos diferentes pero con el mismo nivel de habilidad. El desempeño esperado en un ítem se estima por medio de regresiones no paramétricas ítem – prueba. Si se denota como $E_f(I|M)$ al desempeño esperado en el grupo focal, en el ítem *I*, dada una magnitud de atributo (*M*), y $E_r(I|M)$ al desempeño esperado en el ítem *I* en el grupo de referencia para la misma magnitud de atributo *M*, se puede definir el DIF en el nivel *m* de magnitud de atributo como *Dm*

$= E_{fm} - E_{rm}$. Es decir, la diferencia en el desempeño en un ítem entre individuos del grupo focal y de referencia, habiendo ajustado por la magnitud del atributo que mide la prueba. Tales valores son sometidos inicialmente a una inspección visual, para lo cual se realizan gráficos de dispersión de estas diferencias contra los puntajes totales. En la figura 2 se presentan dos gráficos, el de la 2a corresponde a una situación en la que el DIF puede considerarse despreciable mientras que el 2b muestra una situación de DIF evidente.

La anterior herramienta visual se complementa con el cálculo de un índice que señala los ítems sospechosos de DIF. Este índice se denomina *Diferencia P estandarizada* y se define como:

$$DIFPSTD = \frac{\sum_m w_m (P_{fm} - P_{rm})}{\sum_m w_m} = \frac{\sum_m w_m D_m}{\sum_m w_m}$$

donde P_{fm} y P_{rm} son la proporción de personas que responde correctamente al ítem en cada uno de los grupos (focal y referencia, respectivamente) en el nivel de habilidad m. El término w_m / w_m es un factor de ponderación que se aplica tanto a P_{fm} como a P_{rm} y su elección depende del propósito de la investigación. Algunos valores que puede tomar w_m son: a) el número de examinados en total dentro del estrato m, b) el número de examinados en el estrato m del grupo de referencia, c) el número de examinados en el estrato m del grupo focal o d) la frecuencia relativa en el estrato m en algún grupo de referencia. Los valores que puede tomar el *DIFPSTD* van entre *-1* y *+1*. Si los valores son positivos el ítem está favoreciendo al grupo focal, valores entre *-0.05* y *0.05* se consideran despreciables. Valores entre *0.05* y *0.1* en valor absoluto son dudosos y valores por fuera de estos últimos rangos obligan a una revisión cuidadosa de los ítems en busca de DIF (Dorans y Holland, 1993).

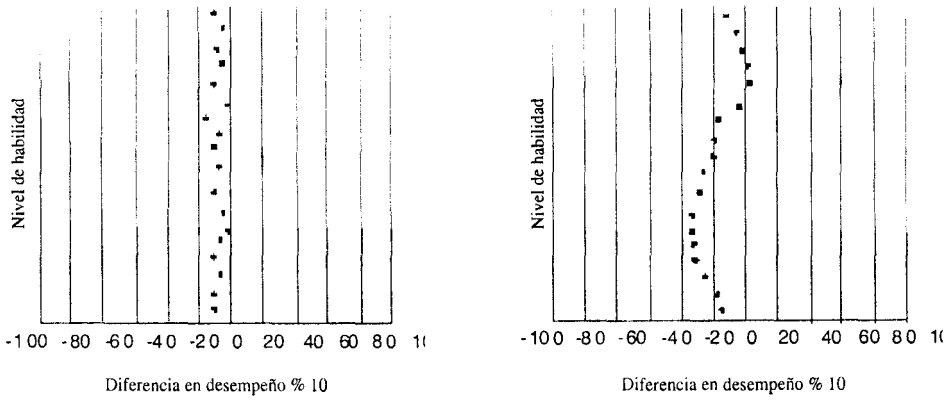


Figura 2. Diferencia en desempeño entre los grupo, en función del nivel de habilidad. (2a) diferencias despreciables, item sin DIF. (2b) diferencias apreciables, ítem con DIF.

Como se mencionó al iniciar este aparte, los tres métodos descritos antes basan la decisión de señalar un ítem con DIF, en la prueba de hipótesis sobre igualdad de proporciones. Los siguientes dos métodos que han sido incluidos dentro de la categoría de los basados en el análisis de tablas de contingencia, se fundamentan en la estimación de los parámetros de un modelo que explique la probabilidad de acierto en el ítem, en función del grupo de pertenencia, de la magnitud de atributo medido y de las interacciones. Rechazar la hipótesis para cada uno de los parámetros, brinda información no solamente sobre la presencia de DIF sino sobre el tipo de DIF.

Los *Modelos Log-lineales* permiten analizar la interacción entre todas las variables simultáneamente, incluyendo además las interacciones entre grupos de variables. Para esta metodología se busca predecir la frecuencia esperada en cada celda de una tabla de contingencia como un producto de efectos (efectos principales e interacciones). Posteriormente se efectúa una transformación logarítmica para convertir dicho producto en un sistema lineal. Cuando no hay ningún efecto, el valor esperado en cada celda es igual al total de observaciones dividido por el número de celdas. Si los totales marginales son diferentes, esto se debe a la presencia de un "efecto" (Powers y Xie, 2000).

Como se sabe, el valor esperado en una de las celdas es $E(X) = N \leftrightarrow p(X)$, donde: N es el número de observaciones en ausencia de efecto. y $p(X)$ es la magnitud de la probabilidad de los diferentes efectos. Bajo la hipótesis nula de independencia de los efectos, la magnitud de la probabilidad de los mismos se expresa en términos multiplicativos, por lo que puede plantearse que el valor esperado en la celda *ai* de la tabla 1, que denominaremos ahora celda (1, 1, 1), es $f_{1,1,1} = N \leftrightarrow \beta_{R1} \leftrightarrow \beta_{G1} \leftrightarrow \beta_{H1}$; donde β_{R1} es el efecto principal del factor R en la celda 1, β_{G1} el efecto principal del factor G en la celda 1 y β_{H1} el efecto principal del factor H en la celda 1. Utilizando transformación logarítmica lo anterior puede expresarse mediante un sistema lineal así:

$$log\lfloor f_{1,1,1} \rfloor = \theta + \lambda_{R1} + \lambda_{G1} + \lambda_{H1}$$

El anterior sistema lineal expresa un modelo de efectos principales. Con base en este sistema también pueden generarse modelos de interacciones dos a dos o de interacciones entre tres efectos. Un modelo con interacciones dos a dos tendría la siguiente estructura:

$$log\lfloor f_{1,1,1} \rfloor = \theta + \lambda_{R1} + \lambda_{G1} + \lambda_{H1} + \lambda_{RG11} + \lambda_{RH11} + \lambda_{GH11}$$

Siguiendo un procedimiento aplicado en las técnicas de regresión, se hace un análisis jerarquizado, empezando por el modelo de efectos principales y terminando en el modelo saturado (modelo que incluye todos los efectos principales y los términos de interacción). Del mismo modo que ocurre en regresión, al introducir términos nuevos cambian los valores de las estimaciones de los parámetros, por lo que se coloca una restricción consistente en que la suma de todos los valores λ de un mismo efecto debe ser cero. En cada etapa del análisis jerarquizado se calcula un valor X^2 con base en las diferencias existentes entre las frecuencias observadas y las esperadas a partir del modelo.

Después de identificar los términos que resulten significativos, éstos se analizan con estadísticos X^2 para ubicar las modalidades de efectos principales o de interacciones que resultan significativos. En la exploración de

DIF con esta metodología los efectos principales están representados por los siguientes componentes: R: Efecto principal dado por respuestas acertadas en el ítem, G: Efecto principal dado por el grupo (de referencia o focal) y H: Efecto principal dado por el nivel de habilidad. Cuando resulta significativo el término de interacción entre los efectos G y R (λ_{GR}), se declara DIF uniforme. Si resulta significativo el término de interacción entre los efectos G, R y H (λ_{GRH}) se considera que existe DIF no uniforme.

Finalmente, la *regresión logística* consiste en la estimación de los parámetros tanto de los efectos de Habilidad (H) y Grupo (G) como de su interacción, en un modelo que explica la probabilidad de acertar al ítem en función de éstos. Dado que la probabilidad p de respuesta correcta a un ítem solo puede tomar valores entre 0 y 1, un modelo lineal de la forma $p = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$ presentaría dos inconvenientes: En primer lugar, el valor de p no puede tomar cualquier cantidad entre más y menos infinito y por otra parte p no sigue una distribución normal sino una Bernoulli o binomial(1,p) (Christensen, 1997).

El primer inconveniente puede resolverse efectuando una transformación logit del valor p. Esta transformación es de la forma

$$logit(p) = ln\left[\frac{p}{1-p}\right]$$

que no es otra cosa que el logaritmo del odds del desenlace de interés. En este caso el modelo sería

$$ln\left[\frac{p}{1-p}\right] = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$$

Dicho modelo puede expresarse como

$$p = \frac{e^{\alpha+\beta_1 x_1+\ldots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\ldots+\beta_k x_k}}$$

Con esta última estructura p toma valores entre 0 y 1.

El segundo inconveniente tiene que ver con el procedimiento de estimación de los parámetros y se resuelve mediante el uso de métodos de máxima verosimilitud. El aporte de las distintas variables que se incluyen en un modelo se evalúa mediante razones de verosimilitud (LR) o con el estadístico de Wald. Generalmente esta última prueba solo suele aplicarse cuando hay que medir un único parámetro Kleinbaum DG. (1994). Para la aplicación de regresión logística en la evaluación de DIF, el modelo tendría la estructura

$$ln\left[\frac{p(R=1)}{1-p(R=1)}\right] = \alpha + \beta_1 H + \beta_2 G + \beta_3 (HG)$$

donde $p(R=1)$ es la probabilidad de que el ítem se responda acertadamente, H el nivel de habilidad y G el grupo (focal o de referencia). Si el coeficiente β_3 es significativamente diferente de cero, se declara DIF no uniforme. Si β_3 es igual a cero y b_2 es significativamente diferente de cero se considera que existe DIF uniforme. Si ninguno de los dos es significativamente diferente de 0, la probabilidad de acertar en el ítem solo depende del nivel de habilidad y no existe DIF.

En los métodos presentados antes de la regresión logística, la magnitud de atributo debía manejarse como una variable categórica. El método de regresión logística tiene la ventaja sobre los previamente anotados, de que considera la naturaleza continua de la magnitud de atributo. Sin embargo, para

algunos autores es un método débil para detectar DIF estrictamente uniforme (Swaminathan y Rogers, 1990).

TÉCNICAS BASADAS EN TEORÍA DE RESPUESTA AL ÍTEM (TRI)

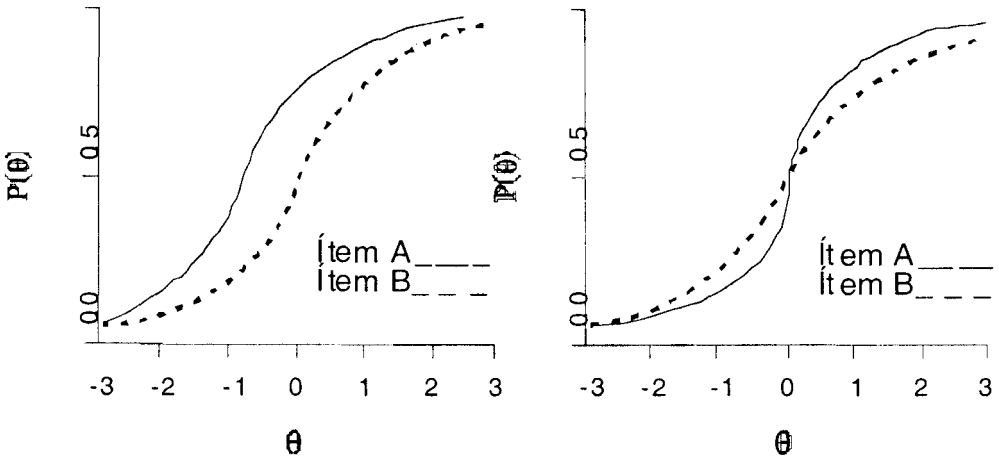
La TRI es un enfoque que se basa en las propiedades de los ítems más que en las de la prueba global (Muñiz, 1997). Los modelos de TRI asumen que existe una relación funcional entre los valores de la variable que mide el ítem y la probabilidad de acertar en el mismo, dicha relación puede expresarse mediante la curva característica del ítem que relaciona la probabilidad de acertar en el mismo, con la magnitud en el atributo que mide la prueba. Si se denota como θ a la habilidad medida por la prueba y como $P(\theta)$ a la probabilidad de respuesta correcta en el ítem para ese θ, la función toma la forma de las curvas presentadas en la figura 1. Como es bien sabido, hay diferentes tipos de modelos TRI dependiendo de los parámetros que se introduzcan; en general, se pueden definir cuatro parámetros: El primero de ellos, θ, se refiere a la magnitud de atributo o habilidad del examinado. Se debe inferir o estimar a partir de las respuestas del examinado, por lo cual se ha denominado habilidad latente. Al ser un valor estimado, θ se mide en una escala arbitraria. Los restantes tres parámetros hacen referencia a las características de los ítems: el parámetro a se refiere a la capacidad de discriminación del ítem, altos niveles de discriminación en un ítem corresponden a CCI más empinadas. El parámetro b representa la dificultad del ítem y se mide en las mismas unidades que θ. El parámetro c, que se conoce como seudo-azar, representa la probabilidad de acertar el ítem al azar. Los valores que toman los anteriores parámetros generan diferentes tipos de CCI como las que se muestran en las figuras 1 y 3.

Los métodos basados en la TRI suponen que un ítem no presenta DIF si los modelos del grupo R y F son iguales. En este caso las gráficas deberían coincidir como se mostró en las curvas de la figura 1a. El punto clave es entonces, la identificación del procedimiento que conduzca a una comparación más fina y precisa de los modelos para los dos grupos. Las diferentes propuestas metodológicas pueden agruparse en tres clases: los que comparan áreas de discrepancia entre las CCI, los que comparan los parámetros de los modelos y los que comparan los modelos.

Los *métodos de comparación de áreas* estiman el DIF mediante la medida de las áreas que separan las curvas para los valores de magnitud de atributo (θ); Rudner, Getson y Knight (1980) propusieron un procedimiento para calcular la discrepancia como la suma de las áreas de diferencia entre las curvas –con la misma métrica para los parámetros– calculadas para valores de θ entre -4 y 4, en incrementos pequeños. Así, el índice de discrepancia está dado por

$$A = \int_{\theta=-4}^{\theta=4} \left| P_R(\theta_j) - P_F(\theta_j) \right| \Delta\theta$$

donde $P_R(\theta_j)$ es la probabilidad que tiene un sujeto del grupo de referencia, con magnitud de atributo igual a θ, de acertar en el ítem, $P_F(\theta_j)$ es misma probabilidad para un examinado con la misma magnitud de atributo pero que pertenece al grupo focal y $\Delta\theta$ es el incremento de θ. Aunque es aconsejable que los incrementos de θ sean muy pequeños ($0{,}005$) para obtener estimaciones más precisas, (Muñiz, 1997) en la figura 4



Los dos ítem difieren en el parámetro b. El ítem A es más fácil que el B.

Los dos ítem tienen igual dificultad pero difieren en el parámetro a.

Figura 3. CCI con diferentes parámetros.

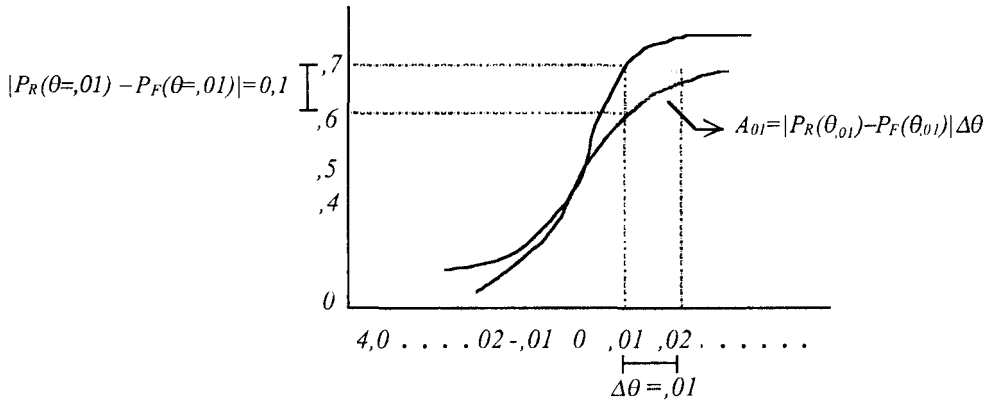


Figura 4. Ilustración del procedimiento de Rudner, Getson y Knight (1980) para la estimación de las áreas de discrepancia entre la CCI de un ítem para dos grupos diferentes.

se ilustra el cálculo del área para $\Delta\theta = 0{,}01$; la zona sombreada sería el valor de discrepancia entre las dos curvas en esos valores de magnitud de atributo y la estimación del DIF, sería la suma de todas las áreas así calculadas, para θ entre -4 y 4. Intuitivamente puede entenderse que un ítem no presenta DIF cuando las áreas son muy pequeñas.

Posteriormente Linn, Levine, Hastings y Wardrop (1981) propusieron diversos índices basados en la idea básica de Rudner, Getson y Knight (1980). Dado que la simple

evaluación del área no tiene en cuenta la cantidad de examinados que se ubican en las diferentes regiones de la superficie evaluada, se han propuesto una serie de ajustes sobre la medición del área, como es el caso de los índices de diferencia de probabilidad de Linn y Harnisch (1981), la suma de cuadrados autoponderada de Shepard, Camilli y Williams (1984), o la medición de áreas con signo de Camilli y Shepard (1994). Aunque estos procedimientos resultan muy sencillos y son frecuentemente utilizados, tienen algunas limitaciones: una de ellas es el tamaño muestral requerido para estimar los modelos TRI y otra es el hecho de que no cuente con un valor crítico para identificar la discrepancia máxima atribuible al azar. Sin embargo, Raju (1990) desarrolló una serie de medidas exactas que permiten hacer prueba de hipótesis mediante el estadístico Z, para modelos TRI de 1, 2 y 3 parámetros (Gómez e Hidalgo, 1997).

Dentro de los *métodos de comparación de los parámetros* de los modelos para los grupos R y F, se destaca el X^2 de Lord. Este método propuesto por Lord (1980) evalúa estadísticamente la diferencia entre los parámetros de los modelos TRI mediante una prueba X^2 que, expresada en notación matricial es: $X^2 = V'S^{-1}V$, donde V el vector de diferencias entre los parámetros estimados para el ítem en los dos grupos y S^{-1} es a inversa de la matriz de varianzas y covarianzas para los vectores de diferencias entre los parámetros. Este estadístico tiene como grados de libertad el número de columnas del vector V, que representa el número de contrastes que se haya efectuado.

Aunque el X^2 de Lord es un procedimiento fácil de utilizar y tiene una prueba de significación, se han señalado algunas limitaciones: no resulta claro su funcionamiento con tamaños de muestra muy pequeños y no se conoce el mínimo tamaño de muestra necesario para que la distribución sea X^2 (Gómez e Hidalgo, 1997); de otra parte, cuando hay impacto entre los grupos, este método puede rechazar como DIF algunos ítems que los métodos de comparación de áreas aceptan como discrepancias pequeñas.

Los *métodos de comparación de modelos* buscan evaluar las diferencias de los modelos IRT para los grupos R y F. Un estadístico de prueba es el G^2, que compara un modelo que incluye parámetros diferentes para los dos grupos (modelo aumentado) con uno en el que los parámetros son iguales (modelo compacto) (Thissen, Stemberg y Gerrard, 1986). La comparación se efectúa con el estadístico: $G^2 = G^2[C] - G^2[A]$, donde C es el modelo compacto y A es el modelo aumentado. Este estadístico sigue una distribución X^2 cuyos grados de libertad son la diferencia en el número de parámetros entre los dos modelos. El rechazo de la hipótesis de igualdad de los modelos, conduce a la conclusión de que el ítem analizado tiene DIF.

En general los métodos basados en la TRI que estiman un nivel de habilidad o atributo latente, gozan hoy de mucha aceptación ya que han mostrado muchas bondades en comparación con otros para la detección de Dif uniforme y no uniforme (Hambleton, Clauser Mazor y Jones, 1993). Sin embargo, presentan la limitación de la exigencia en cuanto a tamaño de muestra para la estimación de los parámetros de los modelos TRI, lo cual dificulta su aplicación en la práctica.

Finalmente, aunque no se ubica en alguna de las categorías de métodos expuestos aquí, vale la pena mencionar brevemente *la*

prueba de sesgo simultáneo de ítems (SSI). Este procedimiento, también conocido como SIBTEST, evalúa simultáneamente el DIF de varios ítems, utilizando un modelo multidimensional, no paramétrico. Tal abordaje permite analizar si un grupo de ítems con DIF tiene un efecto conjunto sobre la puntuación de la prueba generando lo que se denomina funcionamiento diferencial de la prueba (FDP). Este método solo puede aplicarse a pruebas con más de 25 ítems y sin impacto. Se habla de DIF cuando los ítems miden factores diferentes de los que mide la prueba global. Si la prueba pretende medir un nivel de habilidad H, los ítems con DIF, además o en lugar de medir H, medirán habilidades diferentes. De acuerdo con lo anterior puede hablarse de una *subprueba válida*, conformada por los ítems sin sesgo, y de una *subprueba estudiada*, que incluye ítems sesgados. Si se puede determinar la proporción de la prueba que mide la habilidad principal (subprueba válida), es posible emparejar a los miembros de ambos grupos de acuerdo con dicha habilidad.

SELECCIÓN DEL MÉTODO

Dada la variedad y eficiencia de los diversos métodos para evaluación de DIF, se pueden generar interrogantes a la hora de definir cuál de ellos se debe aplicar. No se puede decir que haya un método mejor que otro, independientemente del contexto y las condiciones prácticas de aplicación (Gómez e Hidalgo, 1997; Hambleton, Clauser Mazor y Jones, 1993). La selección del método más apropiado debe supeditarse a una serie de condiciones.

En primer lugar, siempre es preferible un método que iguale los grupos por la magnitud de atributo para no correr el riesgo de confundir DIF con impacto. Desde este punto de vista no resulta recomendable utilizar métodos de la primera categoría (Camilli y Shepard, 1994; Gómez e Hidalgo, 1997). Además, teniendo en cuenta que la detección del DIF en el contexto aplicado se realiza con el fin de tomar decisiones sobre los elementos que componen los instrumentos, es importante considerar si el método dispone de una prueba de significancia que facilite esta decisión.

En segundo lugar, entran en juego otras consideraciones de tipo práctico y del objetivo que se persiga en cada caso. Dentro de las primeras, una de las más importantes es el tamaño de muestra. Si se dispone de grupos grandes - *n>500*, según Camilli y Sheharp (1994), o *n>1000* según Gómez e Hidalgo, (1997) - parece preferible utilizar un método basado en la TRI pero si los grupos son pequeños resulta más recomendable el Mantel-Haenzsel que tiene menos exigencia en cuanto al tamaño de los grupos -100, según Hills (1989) y 200 según Hambleton, Clauser Mazor y Jones, (1993)- En este último caso habrá que tener cuidado de tener suficiente número de casos por celda para que el estadístico no se desestabilice. En Hambleton, Clauser Mazor y Jones, (1993) se encuentran algunas recomendaciones muy útiles a la hora de seleccionar el método para la detección del DIF

Otras consideraciones que dependen del interés de la aplicación particular pueden ser las siguientes: Si se quiere detectar DIF, no solo en ítems individuales sino en grupos de ellos, los métodos recomendados son los modelos log-lineales y la prueba de sesgo simultáneo de ítems (SIBTEST). Si es necesario detectar DIF no uniforme no

resultan recomendables la prueba de sesgo simultáneo de ítems y los métodos de estandarización. Si sencillamente se quiere conocer la magnitud y la dirección del DIF, surge una gama más amplia de opciones, ya que los métodos de medición de área basados en TRI, los métodos de estandarización, la prueba de Mantel-Haenszel, la prueba de sesgo simultáneo de ítems y la regresión logística (transformando coeficientes de regresión en valores OR), permiten tal aproximación.

Otro aspecto que adquiere importancia a la hora de seleccionar el método de análisis es la disponibilidad de programas informáticos para ejecutar los procedimientos. Métodos como la prueba de Mantel-Haenszel, los modelos log-lineales y la regresión logística, pueden ser ejecutados por programas estadísticos a los que se tiene fácil acceso como SPSS®, STATA®, STATGRAPHICS® y BMDP® entre otros. Los métodos basados en la TRI, por su parte, pueden emplearse con ayuda del BILOG para la estimación de los parámetros de los modelos.

REFERENCIAS

Anastasi (1974)*Test Psicológicos* (3ra ed.). Madrid: Aguilar.

Anastasi, A. y Urbina, S (1998). *Test Psicológicos*. (7ª ed.) México: Prentice Hall.

Angoff W.H. (1982). Use of difficulty and discrimination indices dor detecting ítem bias. En RA Berk (Ed), *Handbook of Methods for Detecting Test Bias*. Baltimore: John Hopkins University Press.

Angoff W. H. (1993). Perspectives on Differential Item Functioning Methodology. En P. W. Holland y H. Wainer (Eds), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates, Inc.

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

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

Bond, L. (1987). The Golden Rule settlement: A minority perspective. *Educational Measurement: Isuues and Practice*. 6, 18-20.

Camilli, G. y Shepard, L. (1994). *Methods for identiting biased test items*. London: SAGE Publications.

Christensen R. (1997). *Log-Linear Models and Logistic regression*. 2nd ed. New York: Springer.

Cole, N. (1993). History and Development of DIF. En P. W. Holland y H. Wainer (Eds), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates, Inc.

Dorans, N. J. y Holland, P. W. 1993. DIF Detection and Description: Mantel-Haenszel and Standarization. En P. W. Holland y H. Wainer (Eds), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates, Inc.

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

Faggen, J (1987) Golden Rule revisited: Introduction. *Educational Measurement: Isuues and Practice*. 6, 5-8.

Ferreres, D. (1998). *Funcionamiento diferencial de los items de una prueba de aptitud intelectual en función de la lengua familiar y la lengua de escolarización*. València: Universitat de València. Tesis doctoral inédita.

Fidalgo, A.M. (1996). Funcionamiento diferencial de los items. En J. Muñiz (Coord.) *Psicometría*. Madrid: Editorial Universitas, S.A.

Gómez, J., e Hidalgo, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicótomos: una revisión metodológica. *Anuario de Psicología*. 74, 3-32.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. y Jones, R. W. (1993). Advances in the detection of Differentially functioning test

items. *Journal of Psychological Assessment*, 9 (1), 1-18.

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.

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

Ironson GH. (1982). Use of chi-square and latente trait approaches for detecting item bias. En RA Berk (Ed). *Handbook of Methods for Detecting Test Bias*. Baltmore: John Hopkins University Press.

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

Kahn H.A., Sempos CT. (1989). *Statistical methods in Epidemiology*. New York: Oxford University Press.

Kleinbaum D.G. (1994). *Logistic regression. A self learning text*. New York: Springer.

Lim, R. G. y Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Isuues and Practice*, 6, 13-17.

Lim, R. G. y Drasgow, F. (1990). Evaluation of Two methods for estimating Item Response Theory parameters when assessing differential ítem functioning. *Journal of Applied Psychology*. 75(2), 164-174.

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

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

Lord, F. M.(1980). *Aplications of Item Response Theory to practical testing problems*. Hillsdale: LEA.

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.

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

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

Muñiz, J. (1997) *Introducción a la teoría de respuesta al ítem*. Madrid: Ediciones Pirámide.

Muñiz, J. (1998) *Teoría clásica de los test*. Madrid: Ediciones Pirámide.

Powers, D. A. y Xie, Y. (2000). *Statistical Methods for catagorical Data Analysis*. San Diego: Academic Press.

Prieto Marañón, P.; Barbero García, M. I. y San Luis Costas, C. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and Item Response Theory analysis procedures. *Educational and Psychological Measurement*, 57 (4) 559-568.

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

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. En R. A. Berk, (Ed). *Handbook of methods for detecting test bias*. Baltimore: The Johns Hopkuns University Press.

Rosner B. (1995) *Fundamentals of Biostatistics*. (4a ed). Belmont: Duxbury Press.

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

Scheuneman J.D. (1979). A new method for assesing bias in test items. *Journal of Educational Measurements*:16;143-152.

Selvin S. (1996) *Statistical analysis of epidemiological data*. (2nd ed). New York: Oxford University Press.

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 fundtioning*. Hillsdale, N.J.: LEA.

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

Shepard, L. A. (1982) Definitions of bias. En R. A. Berk (Ed). *Handbook of methods for detecting test bias*. Baltimore: The Johns Hopkuns University Press.

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

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

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

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

Van de Vijver, F. y Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London: SAGE Publications.