Avances en Medición, 6, 5–14
2008

EL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM: UN ASUNTO DE VALIDEZ Y EQUIDAD

Tania Elena Moreira–Mora*
Universidad de Costa Rica, Costa Rica

Resumen

El principal objetivo de este artículo es profundizar en los fundamentos teóricos y metodológicos del funcionamiento diferencial del ítem (DIF), mediante la investigación bibliográfica acerca de algunos conceptos como multidimensionalidad, impacto y funcionamiento diferencial uniforme y no uniforme, y la revisión de algunos procedimientos estadísticos, tanto empíricos como teóricos, para la detección de DIF. Otro tópico destacado es la importancia del análisis de DIF en las pruebas para lograr una medida justa y proporcionar evidencias asociadas a la validez de las inferencias derivadas de las puntuaciones de las pruebas de conocimiento, específicamente al detectar una varianza irrelevante en el constructo medido que provoca diferencias en el comportamiento del ítem. El análisis de DIF permite además obtener una mejor medición de los conocimientos y habilidades y evitar el efecto de las características étnicas, culturales, sociales o de otra naturaleza de los examinados en su desempeño en las pruebas.

Palabras clave: Funcionamiento diferencial del ítem, multidimensionalidad, impacto, métodos empíricos y teóricos, validez y equidad.

Abstract

The main objective of this article is to deepen in the differential item functioning (DIF) theoretical and methodological foundations, using the bibliographical research regarding some concepts like multidimensionality, impact and uniform and non uniform differential functioning; and the revision of some empirical and theoretical statistical procedures for the DIF detection. Another outstanding subject is the DIF analysis importance in the tests to achieve a precise measure and provide evidences associated with the inferences validity, which were derivated from the scores in the knowledge tests; specifically when detecting an irrelevant variance in the measured construct which causes differences in the item performance. The DIF analysis also allows to obtain a better knowledge and abilities measurement and to avoid the effect of the ethnic, cultural, social, or characteristics or from another nature, of those who were examined in their performance in the tests.

Key words: Differential item functioning, multidimensionality, impact, theoretical and empirical methods, validity and equity.

Introducción

Los primeros estudios concernientes al tema del funcionamiento diferencial del ítem (DIF) y de sesgo en las pruebas, tanto de aprovechamiento como de aptitudes, se gestaron en los Estados Unidos de América a partir de los movimientos de los derechos civiles de los años sesenta, como una esperanza hacia el cambio social y educativo. Se buscaba una igualdad de oportunidades dentro de un sistema educativo desigual, lo que se reflejaba en las puntuaciones de las pruebas, usualmente desfavorables a grupos minoritarios tales como negros e hispanos. En este contexto de debates civiles de la sociedad norteamericana sobre la desigualdad de oportunidades entre blancos y las minorías étnicas surgen algunos estudios sobre problemas técnicos en el uso del lenguaje que, por su familiaridad, posibilitaba cierta ventaja para los estudiantes norteamericanos blancos de clase media, desfavoreciendo a los examinados pertenecientes a grupos minoritarios, quienes al no conocer dichos términos tenían rendimientos más bajos, entonces se despierta el interés por la investigación sistemática del sesgo en los ítems (Cole, 1993). También en esta década, apunta Delgado (1998), la crítica feminista fue en gran parte la responsable del

* Apartado postal: 1437-1100 Tibás, San José, Costa Rica, Centroamérica. E-mail: tmoreira@costarricense.cr

impacto del problema del sesgo en la medición, específicamente en la revisión de la sensibilidad de la prueba con la finalidad de eliminar el lenguaje sexista, racista o negativo para cualquier grupo, lo que implicó una representación más balanceada de mujeres y hombres. Esta discusión social sobre la falta de equidad en las pruebas por parte de algunos grupos minoritarios alcanzó las esferas de los pronunciamientos legales y se convirtió en un asunto clave de la investigación psicométrica específicamente en los años 70, con el desarrollo de los métodos para identificar ítems potencialmente sesgados.

En la década de los ochenta los investigadores comienzan a preocuparse por las diferencias observadas en diversos grupos de examinados, especialmente se interesan por buscar explicaciones de la gran disparidad entre los estudiantes "negros e hispanos" con los "blancos" en las pruebas de aprovechamiento. La meta específica de tales estudios fue, y continúa siendo, identificar cualquier ítem con un sesgo contra la minoría y cambiarlo o removerlo de las pruebas para crear pruebas justas y tomar decisiones importantes sobre los estudiantes (Angoff, 1993; Camilli, 1993; Haladyna, 1997; Penfield & Camilli, 2006). Tales preocupaciones evidentemente competen a un asunto de imparcialidad, como lo destaca Andriola (2003): "*la equidad hace referencia a la justicia en la medida, es decir, un test debe posibilitar la igualdad entre los examinados en la probabilidad de acertar a los ítems*" (p. 177). Asimismo, es a mediados de esta década cuando se consolida un marco estadístico general que sirve de soporte para el análisis del funcionamiento diferencial del ítem y se incorpora este análisis como una práctica común en los distintos centros de evaluación educativa de los Estados Unidos, especialmente en las pruebas de selección y estandarizadas. En la década siguiente proliferan las investigaciones relacionadas con este concepto y su medición. En la actualidad es un componente clave en los estudios de validación en virtualmente todas las evaluaciones a gran escala, además de ser uno de los estándares para pruebas psicológicas y educativas (Penfield & Camilli, 2006).

De hecho, el funcionamiento diferencial del ítem es una de las áreas que más investigación e interés político y social ha suscitado en la última mitad del siglo XX en el campo de la medida de variables psicológicas y educativas (Gómez & Hidalgo, 2002; Hidalgo, López & Sánchez, 1997), particularmente por cuatro razones: (a) Las denuncias civiles y legales de algunos grupos minoritarios, como se mencionó anteriormente, (b) el rechazo hacia las pruebas psicométricas en determinados sectores sociales por no representar adecuadamente el grado de ejecución de los distintos subgrupos de examinados (Hidalgo et al., 1997), (c) el amplio uso de las pruebas psicométricas en los procesos de selección, promoción y certificación en los ámbitos educativos y profesionales (Hidalgo, Galindo, Inglés, Campoy & Ortiz, 1999), y (d) la creencia en que el uso de tales pruebas arriesga el futuro de los jóvenes al negarles el diploma, limitar sus posibilidades de obtener empleo o de ingresar a la universidad (Lafourcade, 1994). Estos son los antecedentes del problema del presente artículo concerniente a si la varianza en las puntuaciones se debe a diferencias auténticas en el conocimiento de los examinados, o bien, a fuentes ilegítimas que causan un comportamiento diferencial en los ítems y afectan la validez en la interpretación de los resultados de las pruebas.

Ante esta coyuntura, es una obligación ética y científica garantizar la equidad psicométrica en la medición de los aprendizajes o habilidades. En este sentido afirma Collins (2003) que se deben proporcionar resultados válidos, confiables, sensibles, imparciales y completos para asegurar la medición de los conceptos o conductas que se quieren medir. Por tanto, el principal objetivo de este artículo es profundizar en los fundamentos teóricos y metodológicos del funcionamiento diferencial del ítem y su relación con la validez de las inferencias derivadas de los resultados de las pruebas, sus usos e implicaciones.

Funcionamiento diferencial del ítem

En la actualidad prácticamente ha dejado de utilizarse el término de sesgo en favor del término *Funcionamiento diferencial del ítem,* acuñado por Holland & Thayer (1988) a finales de la década de los

ochenta (FDI sigla en Español, y DIF sigla en inglés proveniente de Differential Item Functioning). Una razón ha sido la proliferación en los últimos años de técnicas para la detección del DIF y la otra, la potencia y fuerza de estas técnicas para detectar únicamente el DIF, sin permitir encontrar posibles explicaciones acerca de su naturaleza o causa (Andriola, 2001). Igualmente, destacan Cahalan y Cook (s.f.) que el estudio del comportamiento diferencial de los ítems ha sido una forma de aproximarse a la problemática del sesgo en la medición de los aprendizajes, pues es un indicador de las diferencias en la ejecución de un ítem entre miembros comparables -con la misma puntuación en una evaluación- de grupos diferentes. En este sentido, Hidalgo et al. (1999) ratifican que una forma de investigar las razones por las cuales un grupo es sistemáticamente mejor o peor que otro es atender a las características de los ítems; esto significa que uno o más ítems no funcionan adecuadamente, de forma que tiende a favorecer a un grupo (referencia) sobre otro (focal).

Desde el punto de vista de la psicometría, un ítem presenta funcionamiento diferencial si sujetos con un nivel idéntico respecto al atributo medido en la prueba y pertenecientes a distintas subpoblaciones o grupos culturales no tienen la misma probabilidad de responder correctamente el ítem o la prueba (Anastasi & Urbina, 1998; Atorressi, Galibert, Zanelli, Lozzia & Aguerri, 2003; Camilli, 1993; Hidalgo et al., 1999; Muñiz, 1990; Padilla, González & Pérez, 1998; Penfield & Camilli, 2006).

Un aspecto por aclarar es que un ítem con un funcionamiento diferencial no implica necesariamente sesgo o una condición de injusto (Montero, 1993; Penfield & Camilli, 2006). Como lo aclara Montero (1993), cuando el DIF resulta de la medición de atributos irrelevantes entonces es una evidencia de sesgo; mientras que existe DIF legítimo cuando es producto de la variabilidad en el constructo relevante que se desea medir; es decir, cuando representa diferencias en el conocimiento o habilidad entre los miembros de los grupos examinados. En síntesis, el DIF es un análisis que atañe a las evidencias asociadas a la validez de la interpretación de las puntuaciones de las pruebas de conocimiento, específicamente, con la varianza irrelevante en el constructo medido que provoca diferencias en el comportamiento del ítem. Como lo afirma Elosua (2003), la detección estadística del DIF no es un fin en sí mismo, es un instrumento útil que adquiere relevancia dentro de un marco sustantivo de estudio de la validez.

Asimismo, Schmitt, Holland y Dorans (1993) agregan que es necesario identificar los factores relacionados con el DIF y complementarlos con una teoría de la dificultad diferencial del ítem, la cual no ha sido suficientemente desarrollada, y con una evaluación sistemática de hipótesis de DIF, compleja por sus consideraciones prácticas y éticas. Según los autores, estas hipótesis pueden generarse a partir de consideraciones teóricas o empíricas. Las hipótesis teóricas del DIF se fundamentan en los conocimientos previos pertinentes a los procesos cognitivos que podrían estar relacionados con el desempeño diferencial en los ítems en cualquier subpoblación de examinados, mientras que las hipótesis empíricas, generadas después del análisis DIF, pueden sugerir que ciertas características de los ítems están relacionadas diferencialmente en uno o más subgrupos de la población. Adicionalmente, como apuntan Penfield y Camilli (2006), es necesaria una evaluación sistemática de las hipótesis de DIF, cuyo primer paso supone una medición de DIF para usar esa información en la generación de hipótesis y, posteriormente, una evaluación confirmatoria de esas hipótesis, mediante un estudio aleatorio y controlado del DIF con ítems construidos para probar tales hipótesis. Si en las muestras aleatorias de examinados se exhibe diferentes niveles de DIF en los ítems estudiados, entonces existe evidencia que la propiedad del ítem manipulada es causa de DIF.

Posiblemente las explicaciones teóricas se encuentren en el contexto de los grupos poblacionales de interés al informar sobre el comportamiento de las preguntas, como una estrategia para reconocer la posible influencia del contenido curricular u otros aspectos de la prueba (como el formato o el lenguaje) en las respuestas de los examinados y sus efectos sobre la medida del constructo de interés cuando sus miembros tienen una menor probabilidad de tener éxito, en comparación con sujetos de otro grupo igualmente competentes en el constructo medido; por tanto, es importante contemplar la posibilidad de realizar un estudio que incorpore la interacción cultura – evaluación, así como el distinto bagaje social y

económico de los sujetos, que ofrezca una aclaración de las causas del DIF (Delgado, 1998; Hidalgo et al., 1999; Muñiz, 1990; Montero, 1993; Pardo, 1999).

Impacto

Normalmente la comparación se realiza entre el grupo de análisis e interés principal llamado focal y el grupo que sirve como base de comparación denominado de referencia (Donoghue, Holland & Thayer, 1993; Hidalgo et al., 1999; Montero, 1993). Un requisito fundamental que debe cumplir cualquier procedimiento estadístico en la comparación de estos dos grupos es no confundir las diferencias reales entre los grupos con respecto a la variable o constructo medido, lo que técnicamente se llama *impacto*, con las diferencias provocadas por un funcionamiento diferencial de los ítems (Andriola, 2002; Fidalgo, Mellenbergh & Muñiz, 1999; Hidalgo et al., 1999). A modo de ilustración, un problema aritmético puede medir habilidad matemática en un grupo y en otro la habilidad de comprensión verbal. En tal caso existe un DIF causado por una varianza irrelevante en el constructo medido; no obstante, si se encuentra que un ítem con decimales resultó más difícil para cierto grupo, esta diferencia sí es relevante para el constructo de la habilidad matemática, por consiguiente, tal variación es resultado de las diferencias reales entre los grupos de examinados. En este sentido, Dorans y Holland (1993) aclaran que el impacto se refiere a las diferencias en el desarrollo de la habilidad entre dos grupos intactos, tales como aquellos definidos por los rasgos étnicos o el género; por consiguiente, las diferencias en la ejecución de un ítem refleja las diferencias en la distribución de la habilidad promedio de cada grupo.

La diferenciación entre ambos conceptos es fundamental para los constructores de prueba, puesto que les ayuda a demostrar si el funcionamiento diferenciado de la prueba es debido al DIF o al impacto. En caso de DIF, los estudios se pueden centrar en aquellas características del ítem o de la prueba que sean irrelevantes al constructo; o bien, en los posibles factores cognoscitivos relevantes en la medición del constructo (Taylor & Lee, 2008).

Multidimensionalidad

En la actualidad se examina el criterio de la multidimensionalidad como una posible causa del DIF que implica reconocer la presencia de uno o más factores secundarios no intencionados, además del principal medido por los ítems en uno de los grupos comparados y que no sea relevante al propósito por el cual fue construida la prueba (Andriola, 2002; Angoff, 1993; Middleton, 2008; Penfield & Camilli, 2006). En tal condición puede suceder que un ítem presente un DIF por medir un constructo secundario en uno de los dos grupos comparados, por ejemplo comprensión de lectura (factor secundario) en un ítem construido para medir la habilidad de sumar (factor principal). Entonces, claramente, este funcionamiento diferencial se presenta cuando no se satisface el supuesto de unidimensionalidad (Lozzia, Galibert, Aguerri & Attorresi, 2005).

Debido a esta relación entre el DIF con la multidimensionalidad se han llegado a considerar como sinónimos, por lo que Zieky (1993) aclara esta confusión tomando como ejemplo una prueba de matemática en la que la mayoría de los ítems miden conocimientos de álgebra y se le incluyen unos pocos ítems de geometría; entonces es de esperar que estos ítems resulten con un elevado valor de DIF. No obstante, si la prueba midiera principalmente geometría con unos pocos ítems de álgebra, entonces, de igual modo, serían identificados con un DIF. Además, aclara que los métodos basados en la Teoría de Respuesta al Ítem, así como el Mantel-Haenszel y el método de estandarización (métodos empíricos) identificarían con DIF los ítems que no miden la misma dimensión o dimensiones que la mayoría de la prueba o escala. Finalmente, apunta Zieky que la imparcialidad de un ítem depende directamente del propósito por el cual fue utilizado en la prueba y su funcionamiento diferencial es un asunto extremadamente complejo, que toca algunos puntos muy sensibles.

Funcionamiento diferencial del ítem uniforme y no uniforme

Desde los supuestos de la Teoría de Respuesta al Ítem (TRI), un constructo o variable latente (θ), no observable directamente, se puede estimar para cada sujeto a partir de las respuestas dadas en un instrumento de medición con base en tres modelos psicométricos. El principal, denominado Rasch, estima únicamente el valor del parámetro *b* (dificultad). En el modelo de dos parámetros se consideran *b* y *a* (la discriminación), y en el modelo de tres parámetros debe estimarse también el valor de *c* o del azar (Kohan, 2004; Montero, 2001; Muñiz, 1990; Penfield & Camilli, 2006). Entonces, un ítem presenta un funcionamiento diferencial cuando la probabilidad de ser resuelto correctamente por los examinados con el mismo nivel de habilidad varía en función de su grupo de pertenencia (sexo, cultura, nivel socioeconómico…), en otros términos, cuando las funciones de respuesta o curvas características de los ítems (CCI) son diferentes para distintas poblaciones (Anastasi & Urbina, 1998; Elosúa & López, 1999; Kohan, 2004). La curva característica del ítem es una función matemática que establece la relación existente entre las puntuaciones de los sujetos en la variable medida (θ) y la probabilidad de responder correctamente al ítem (Hidalgo-Montesinos & López-Pina, 2002, Kohan, 2004; Pardo, 1999).

El *funcionamiento diferencial uniforme o consistente* se presenta cuando no existe interacción entre el nivel del atributo medido y la pertenencia a un determinado grupo, es decir, cuando el ítem proporciona una ventaja constante para el mismo grupo de un extremo a otro del rango de desempeño, por tanto, las curvas características del ítem son paralelas (Andriola, 2001; Andriola, 2002; Hidalgo et al., 1999; Koretz, 1997; Penfield & Camilli, 2006). Al observar la figura (1a) la curva característica del ítem (CCI) de las mujeres (grupo focal) está situada más a la izquierda que la de los varones (grupo de referencia), lo que indica que el ítem es más fácil para ellas. La diferencia en el parámetro b (dificultad) supone que el ítem presenta DIF, por ejemplo los varones y mujeres que poseen θ = 0 (valor medio de la distribución) poseen diferentes probabilidades de acierto [P (θ)]: 0,65 los varones y 0,85 las mujeres. Esta diferencia en los valores P (θ) supone que el ítem es favorable a las mujeres, puesto que para un mismo valor de (θ) las probabilidades de responder correctamente es siempre superior para las mujeres (Andriola, 2001).

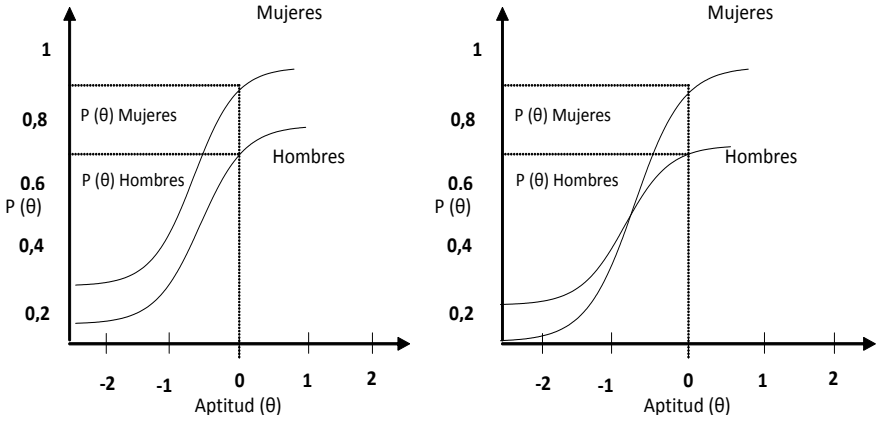

Figura 1 Representación gráfica de dos ítems con DIF para dos grupos diferentes. (a) DIF uniforme. (b) DIF no uniforme. Tomado de Andriola (2001).

El *funcionamiento diferencial no uniforme o inconsistente* se observa cuando la diferencia en las probabilidades de responder correctamente al ítem no es la misma a lo largo del continuum del atributo medido; es decir, cuando aparece un DIF en un nivel específico, por ejemplo el grupo de referencia puede tener una pequeña ventaja en niveles bajos de θ, pero en los niveles altos la ventaja es mayor entre ambos grupos (Andriola, 2001, 2002; Hidalgo et al., 1999; Koretz, 1997; Penfield & Camilli, 2006). Dentro del DIF no uniforme existe una modalidad conocida como cruzamiento (crossing) cuando el grupo de referencia tiene una relativa ventaja en uno de los extremos del continuum de θ, mientras que el grupo focal muestra una cierta ventaja en el otro extremo del continuum, como se ilustra en la Figura 1b, en donde los parámetros *a*, *b* y *c* tienen valores distintos en las CCI, es

decir, no son paralelas. Esta distinción entre DIF uniforme y no uniforme es de crítica importancia cuando se analizan e interpretan las estadísticas del comportamiento diferencial (Penfield & Camilli, 2006).

Métodos para la detección del DIF

Desde la década de los setenta han surgido múltiples procedimientos estadísticos para la detección del DIF en pruebas de conocimiento construidas con ítems dicotómicos y politómicos para superar algunas limitaciones metodológicas y lograr una mayor convergencia de los resultados. Estos procedimientos se han llegado a clasificar, según el modelo psicométrico, en *empíricos* y *teóricos*.

Los métodos empíricos, también conocidos como métodos de invariancia condicional observada o métodos condicionales, se fundamentan en las puntuaciones observadas en la prueba desde la perspectiva de la Teoría Clásica de los Tests, tales como el método del Delta Gráfico definido por William H. Angoff en 1972, el chi-cuadrado de Scheuneman del año 1979, el estadístico de Camilli en 1979, el método logístico iterativo propuesto por F.B. Baker en 1981, el chi-cuadrado de Pearson o total, el método de estandarización planteado por Dorans y Kulick en 1986, el Mantel-Haenszel desarrollado por Holland y Thayer en 1988, el método de regresión logística propuesto por Rogers y Swaminatan en 1990, el análisis discriminante logístico de Miller y Spray en 1993, y las variaciones iterativas sobre algunos de ellos (Andriola, 2002, 2003; Hidalgo et al., 1997; Montero, 1993; Wainer, 1993). Estos métodos empíricos o condicionales están basados en la conocida Paradoja de Simpson, que de acuerdo con Dorans y Holland (1993) "*enfatiza la importancia de comparar lo comparable*" (p. 38); es decir, se resalta la importancia de comparar la probabilidad de acierto para un determinado ítem, teniendo siempre en cuenta que los sujetos comparados sean del mismo nivel de habilidad en el constructo medido. Una limitación al utilizar sólo las frecuencias observadas de acierto en los grupos comparados es que los resultados se pueden afectar por la presencia de diferencias reales entre los grupos o impacto (Andriola, 2003).

Los métodos teóricos, también conocidos como métodos de invariancia condicional no observada o métodos incondicionales, se centran en modelos matemáticos como los de la Teoría de Respuesta a los Ítems por utilizar las estimaciones de la habilidad (θ), según el modelo más adecuado a los datos (Andriola, 2002, 2003; Hidalgo et al., 1997; López, Pérez & Armendáriz, 2005; Montero, 1993; Wainer, 1993). Específicamente, los procedimientos más utilizados para verificar el DIF han sido la comparación de parámetros, al que corresponde la prueba normal para la diferencia de los parámetros de dificultad, el análisis del área limitada por dos CCI en las poblaciones de referencia y focal, la comparación del área exacta con signo y del área exacta sin signo de Raju, el método del Chi-cuadrado de Lord, la comparación del ajuste de los modelos, concretamente la prueba de la razón de verosimilitud o el método de las probabilidades desarrollado por Camilli y Shepard en 1994, y las propuestas iterativas sobre algunos de estos estadísticos (Atorressi et al., 2003; Andriola, 2002, 2003; Elosúa & López, 1999; Hidalgo et al., 1997). Estos modelos de la TRI son muy utilizados por ofrecer un marco teórico adecuado para el análisis a través del estudio de las curvas características de los ítems (CCI) o funciones de respuesta a los ítems (FRI).

En cuanto a las principales debilidades metodológicas de estos procedimientos se puede señalar los riesgos potenciales de cometer Error tipo I (falsos positivos) o Error de tipo II (no detectar un ítem con DIF), los efectos del tamaño de la muestra en la identificación del DIF y la falta de prueba de hipótesis. Para contrarrestar tales debilidades se recomienda aplicar el procedimiento seleccionado en dos o más etapas (Marco, 1977, citado en Lord, 1980; Segal, 1983, citado en Candell & Drasgow, 1988; Miller & Oshima, 1992; Hidalgo-Montesinos & López-Pina, 2002), utilizar varias técnicas de identificación del DIF como un criterio externo de validación y de estabilidad; o bien, recurrir a diferentes tamaños de muestras y de habilidades de los sujetos (Aguerri, Galibert, Zanelli & Attorresi, 2005).

Conclusiones derivadas de múltiples investigaciones coinciden, además, en la necesidad de conjugar estos métodos estadísticos, cada vez más sofisticados, con la opinión de expertos en el área temática y la instrucción, con el objetivo de identificar los constructos irrelevantes no intencionados medidos por los

ítems que podrían ser potenciales fuentes de DIF, explicar las razones por las cuales un ítem funciona de manera distinta (favorable o desfavorable) en un determinado grupo de examinados y así ayudar a los constructores de pruebas a comprender aquellas propiedades del ítem y del grupo que son responsables del desempeño diferencial en la prueba y, en consecuencia, conducir a pruebas que sean más justas (Angoff, 1993; Attorresi et al., 2003; Andriola, 2002, 2003; Camilli, 1993; Elosúa & López, 1999; Hidalgo et al., 1999; Montero, 1993; Penfield & Camilli, 2006). Usualmente uno de los procedimientos es establecer una correlación entre la clasificación predicha por los jueces con las mediciones empíricas del DIF, mientras que otra estrategia ha sido que los expertos expliquen (en vez de predecir) los patrones de los ítems sesgados hasta llegar a un acuerdo satisfactorio (Camilli, 1993; Montero, 1993).

Finalmente, cabe señalar que algunos de los rasgos comunes tanto de los procedimientos empíricos como teóricos consisten en cumplir con el supuesto de *unidimensionalidad* de la TRI, que de acuerdo con Muñiz (1997), radica en que los ítems constituyen una sola dimensión, es decir, miden un mismo y único rasgo; además los índices de DIF se basan en un criterio interno, la puntuación total de la prueba, por tanto no pueden ir más allá de la prueba (circularidad interna) para evaluar el impacto en otras formas de medida del mismo constructo (Angoff, 1993).

Validez

El funcionamiento diferencial de los ítems es un análisis que concierne a las evidencias de validez del significado y de las interpretaciones derivadas de los resultados. En este sentido, Valverde (2001) subraya que la validez no es una propiedad intrínseca de las pruebas, sino una propiedad de las interpretaciones y los usos que se propone dar a los datos obtenidos. Además, en la actualidad converge la noción tradicional de validez con el concepto unificado propuesto por Samuel Messick. La noción convencional se relaciona con lo que se mide, qué tan bien lo hace y con las inferencias de sus resultados (Anastasi & Urbina, 1998; Kerlinger, 1994; Mehrens & Lehmann, 1982). Desde el punto de vista psicométrico, se han dado diversas definiciones de la validez referentes al criterio, contenido y constructo.

Las evidencias de validez de criterio o predictiva se refieren al uso del instrumento para estimar alguna conducta criterio externa al instrumento de medición (Nunnally & Bernstein, 1995). Las asociadas al contenido se basan en una medida del grado en que una prueba representa o muestrea el universo de ítems de una determinada área de conocimientos, con base en el criterio de jueces. La validez de constructo es el grado en que mide un rasgo teórico y explica las diferencias individuales en las puntuaciones de las pruebas. No se trata sólo de validar una prueba, también se debe validar la teoría que fundamenta la prueba (Anastasi & Urbina, 1998; Kerlinger, 1994). Los constructos, según Anastasi y Urbina (1998) "*son categorías amplias que se derivan de los rasgos comunes que comparten las variables conductuales observables directamente; pero se trata de entidades teóricas que por su parte no pueden ser observadas de manera directa*" (p. 114). No obstante, esta perspectiva tradicional de la validez en las pruebas de conocimiento, habilidad y de aptitud fue cuestionada principalmente por Samuel Messick, quien propuso un concepto unitario al considerar que sólo existen diferentes formas para encontrar evidencias de validez.

En esta investigación teórica se trascienden las diferencias entre los distintos procedimientos de validación para asumir la concepción unificada planteada por Samuel Messick, quien propone que la validez encierra "*…un juicio integrado y evaluativo del grado en que la evidencia empírica y las razones teóricas apoyan lo adecuado y lo apropiado de las interpretaciones y las acciones basadas en las puntuaciones de las pruebas u otras formas de evaluación*" (Messick, 1995, p. 5).

Es más, conforme con los estándares para las pruebas educativas y psicológicas establecidos por la American Educational Research Association (AERA), la American Psycological Association (APA) y la National Council on Measurement in Education (NCME), el proceso de validación involucra la acumulación de evidencia que proporciona una base científica para la interpretación de las puntuaciones de la prueba y la relevancia (1999).

Asimismo, para Messick (1989) no es suficiente ni la validación de contenido, ni la de criterio. En el caso de la primera, basada en juicios de expertos sobre la relevancia del contenido de un dominio conductista específico y lo representativo de cada ítem en cuanto a cómo abarca ese dominio, no se preocupa por los procesos de respuesta, las estructuras internas o externas de las pruebas, las diferencias de desempeño entre los grupos y contextos. Efectivamente, aporta evidencia de apoyo con respecto a la relevancia y representatividad del dominio del contenido; mas no proporciona evidencia que sustente las inferencias extraídas de las notas de las pruebas. En contraste con la validación de contenido, la relacionada con los criterios está respaldada en el grado de correlación empírica entre las notas de la prueba y las puntuaciones de criterio, y, como tal, depende de elementos seleccionados de la estructura externa de la prueba. Consecuentemente, no se preocupa por ningún otro tipo de evidencia, excepto las correlaciones específicas de criterio de la prueba, las cuales podrían ser potencialmente deficientes al capturar el dominio del criterio de interés, ser contaminadas por variaciones irrelevantes y sujetas a error. Por ello, es necesario valorar las medidas de criterio y las pruebas en relación con las teorías de constructo del dominio del criterio.

En cuanto a la validación del constructo existen dos amenazas principales, una es la subrepresentación de la prueba al ser demasiado estrecha y no incorporar ciertas dimensiones o facetas importantes del constructo. Esto involucra un significado reducido de las puntuaciones porque las pruebas no muestrean adecuadamente algunos contenidos, procesos psicológicos o elimina algunas formas de respuesta comprendidas en el constructo, como por ejemplo, una prueba de comprensión de lectura para medir en los niños la habilidad de lectura y de interpretación de la historia puede estar subrepresentada al no contener una variedad suficiente de pasajes o al ignorar un material de lectura común. La otra es la variación irrelevante referida al grado en que las puntuaciones de la prueba son afectadas por procesos extraños al constructo o al contener un exceso de variación, permitiendo que los ítems sean más fáciles o más difíciles para ciertos examinados. En el caso de la prueba de comprensión de lectura algunos componentes irrelevantes podrían ser reacciones emocionales a los contenidos de la prueba, familiaridad con los contenidos de los pasajes de lectura, conocimiento del vocabulario o la rapidez de la lectura (AERA, APA, NCME, 1999; Messick, 1989). Precisamente, el funcionamiento diferencial de los ítems corresponde a esta segunda amenaza, al detectar variancias irrelevantes del constructo medido en la prueba, como por ejemplo comprensión de lectura en un ítem de razonamiento matemático.

Por ello, destaca Montero (1993), es necesario analizar si las fuentes del funcionamiento diferencial son externas como por ejemplo las deficiencias instruccionales, o por el contrario, son producto de una variancia ilegítima en el constructo medido, en tal caso es preciso alertar a los constructores de pruebas para que cambien o modifiquen los ítems sesgados y así aumentar la validez de las puntuaciones, con el fin de lograr una mayor exactitud en la medición del constructo al eliminar los impedimentos irrelevantes provocados, entre otras fuentes, por el formato, el léxico, los contenidos, las imágenes o la redacción del encabezado y opciones del ítem. Con este análisis se busca demostrar cuán adecuadas son las interpretaciones, usos y consecuencias sociales de las puntuaciones, que dependen no solo de las características de los ítems, sino también del nivel cognoscitivo de los examinados y del contexto evaluativo.

Conclusiones

El propósito cardinal de esta investigación fue profundizar tanto en los principios teóricos como metodológicos del funcionamiento diferencial de los ítems, debido a la relevancia psicométrica de este análisis en las pruebas de selección y estandarizadas, particularmente, para demostrar una equidad en la medición, obtener resultados con más evidencias de validez, controlar a priori el DIF y generar teorías sobre las fuentes de variancia irrelevante del constructo medido en las pruebas. Conforme con los hallazgos de esta investigación se puede afirmar que la vertiente más desarrollada en este campo ha sido la metodológica, pues los estudios se han enfocado principalmente en los múltiples métodos estadísticos para detectar el DIF en pruebas de conocimiento y de aptitudes. En concreto, se han preocupado por obtener

mejores estimaciones de los valores del DIF, tanto en los métodos empíricos como teóricos, por detectar el DIF uniforme y no uniforme, por las pruebas de significación estadística y por minimizar el riesgo de incrementar el error tipo I (eliminar ítems que no tienen DIF) o el error tipo II (mantener o no revisar ítems con DIF) y por la estabilidad de los resultados del DIF.

En la vertiente teórica se han logrado importantes hallazgos centrados en la identificación de algunas fuentes del DIF. En general, los estudios se han realizado comparando el género y la etnia de los grupos examinados, también en menor medida diferencias lingüísticas y culturales. Sin embargo, estas investigaciones no han sido suficientemente desarrolladas por la complejidad de generar explicaciones sustantivas acerca del DIF en ciertos grupos minoritarios o con desventaja social. El desarrollo de tales teorías implica el planteamiento y comprobación de hipótesis, lo que requiere de una evaluación sistemática, considerando tanto las evidencias estadísticas como los conocimientos previos relacionados con el DIF en el grupo de interés mediante diseños experimentales, especialmente con ítems construidos para probarlas.

A pesar de las limitaciones de algunos procedimientos estadísticos y de la falta de comprobación de hipótesis, el análisis del DIF es elemental para incrementar las evidencias de validez en las inferencias de las pruebas de selección y estandarizadas. Ahora bien, al asumir la validez como un concepto unificado para apoyar lo adecuado de las interpretaciones, usos y consecuencias sociales de las puntuaciones de las pruebas, se busca comprobar que las puntuaciones obtenidas por los examinados estén en función tanto de las características de los ítems como de los estudiantes y del contexto evaluativo, de esta forma se trascienden las evidencias limitadas de contenido, de criterio o de constructo. Asimismo, el análisis del DIF es un asunto de equidad en la medición de los aprendizajes y de las aptitudes, pues todos los estudiantes deben tener las mismas condiciones para demostrar sus conocimientos y habilidades en el constructo medido eliminando cualquier barrera de formato, lenguaje, contenido, distractores o cualquier otra fuente que proporcione una ventaja inapropiada a un determinado grupo de examinados.

Igualmente es importante subrayar que el estudio del DIF se debe complementar con otros análisis estadísticos como el de la estructura factorial para descartar como una posible fuente la multidimensionalidad del ítem; el de confiabilidad para garantizar la consistencia interna y la estabilidad de las puntuaciones, también con la técnica de los jueces para explorar las posibles fuentes por las que un ítem funciona distinto en un grupo de examinados y determinar si esas diferencias son legítimas del constructo medido, o por el contrario, son causadas por una variancia irrelevante de ese constructo.

En síntesis, el análisis del DIF concierne a dos asuntos álgidos, equidad y validez de la interpretación de los resultados, sus usos e implicaciones, al comprobarse que es una técnica útil para evidenciar el grado de validez y de justicia en la medida. Los involucrados en la construcción de pruebas estandarizadas, por tanto, deben incorporar este análisis como una práctica usual en los estudios estadísticos de los resultados con la finalidad de lograr instrumentos invariantes y una mejor estimación del constructo en los diferentes grupos de examinados. Igualmente, una tarea imprescindible en el ámbito investigativo es indagar en las posibles fuentes de DIF en diferentes subgrupos de examinados para empezar a desarrollar teorías que permitan comprender e interpretar el comportamiento diferencial de los ítems.

Referencias

Aguerri, M.E., Galibert, M.S., Zanelli, M.L. & Attorresi, H.F. (2005). Detección errónea del funcionamiento diferencial del ítem. Una comparación de métodos. *Psicothema*, 17, 350-355.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, USA: Autor.

Anastasi, A. & Urbina, S. (1998). *Tests psicológicos* (7ª ed.). México: Prentice Hall.

Andriola, W. B. (2001). Determinación del funcionamiento diferencial de los ítems (DIF) destinados a la evaluación del razonamiento verbal a partir del tipo de escuela. *Bordón.Revista de pedagogía*, *53*(4), 473 – 484.

Andriola, W. B. (2002). *Detección del funcionamiento diferencial del ítem (DIF) en test de rendimiento. Aportaciones teóricas y metodológicas*. Tesis de Doctorado para optar al título de Doctor, Universidad Complutense de Madrid,

España. Recuperado el 22 de octubre del 2005 de: http://tede.ibict.br/tde_arquivos/1/TDE-2004-12-13T07:17:46Z-57/Publico/1 WagnerBandeiraAndriola_intro_cap8.pdf

Andriola, W. B. (2003). Descripción de los principales métodos para detectar el funcionamiento diferencial del ítem (DIF) en el área de la evaluación educativa. *Revista de Pedagogía Bordón*, 55(2), 177 – 189.

Angoff, W. H. (1993). Perspectives on Differential Item Functioning Methodology. En P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 3-23). New Jersey: Lawrence Erlbaum Associates, Inc.

Attorresi, H.F., Galibert, M.S., Zanelli, M., Lozzia, G. & Aguerri, M. E. (2003). Error tipo I en el análisis del funcionamiento diferencial del ítem basado en la diferencia de los parámetros de dificultad. [Versión electrónica]. *Psicológica*, *24*, 289 – 306.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedure obscure test fairness issues? En P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). New Jersey: Lawrence Erlbaum Associates, Inc.

Cahalan, C. & Cook, L. (s.f.). *An examination of differential item functioning for students with disabilities*. Educational Testing Service. Recuperado el 16 de marzo del 2005 de: www.ets.org/research/fellowship/fe100res_projs.html

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

Cole, N. S. (1993). History and development of DIF. En P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 25–29). New Jersey: Lawrence Erlbaum Associates, Inc.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, 12, 229–238.

Kohan, N. (2004). Teoría de Respuesta al ítem: Supuestos básicos. [Versión electrónica]. *Evaluar*, *4*, 95-110.

Delgado, C. (1998). El problema de sesgo en los test. Revisión histórica y cuestiones críticas. *Revista de Ciencias Sociales*, *80*, 21– 44.

Donoghue, J., Holland, P. & Thayer, D. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. En P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 137–166). New Jersey: Lawrence Erlbaum Associates, Inc.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. En P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 35–66). New Jersey:Lawrence Erlbaum Associates, Inc.

Elosua, P. & López, A. (1999). Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. [Versión electrónica]. *Psicológica*, *20*, 23 - 40.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema*, 15, 315-321.

Fidalgo, A. M., Mellenbergh, G. J. & Muñiz, J. (1999). Aplicación en una etapa, dos etapas e iterativamente de los estadísticos Mantel-Haenszel. [Versión electrónica]. *Psicológica*, *20*, 227-242.

Gómez, J. & Hidalgo, M. D. (2003, Julio 24). La validez en los test, escalas y cuestionarios. *La sociología en sus escenarios, 8.* Recuperado el 21 de octubre del 2004 de: http://huitoto.udea.edu.co/~ceo/revista_filtrados.php?id=20&autor=15&otro=102

Haladyna, T. (1997). *Writing test items to evaluate higher order thinking.* MA: Allyn & Bacon.

Hidalgo, M. D., Galindo, F., Inglés, C. J., Campoy, G. & Ortiz, B. (1999). Estudio del funcionamiento diferencial de los ítems en una escala de habilidades sociales para adolescentes. [Versión electrónica]. *Anales de Psicología*, *15*(2), 331-343.

Hidalgo, M. D., & López-Pina, J. A. & Sánchez, J. (1997). Error tipo I y potencia de las pruebas chi-cuadrado en el estudio del funcionamiento diferencial de los ítems. *Revista de Investigación Educativa*, *15*(1), 149 – 170.

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

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

Kerlinger, F. (1994). *Investigación del comportamiento* (2ª ed.). México: McGraw-Hill.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky*. Los Angeles: Center for Research on Evaluation, Standards and Student Testing, UCLA (CSE Technical Report 431). Recuperado el 16 de marzo del 2005 de: http://www.cse.ucla.edu/products/Reports/TECH431.pdf

Lafourcade, P. (1994). La calidad de la educación y el cuestionamiento de las pruebas estandarizadas de rendimiento en Estados Unidos de Norteamérica: Las nuevas búsquedas. *Revista Iberoamericana de Educación 2* (3), 53-61.

López, J., Pérez, T.A. & Armendáriz, A.J. (2005). La evaluación mediante tests: ¿Por qué no usar el ordenador? Revista Iberoamericana de Educación, 36/11. Recuperado el 27 de abril del 2007 de http://www.rieoei.org/deloslectores/1040Lopez.PDF

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

Lozzia, G.S., Galibert, M.S. Aguerri, M.E.& Attorresi, H.F. (2005). Construcción de un banco de ítem de Razonamiento verbal. *Interdisciplinaria Revista de Psicología y Ciencias afines*, 22, 5-27.

Mehrens, W. & Lehmann, I. (1982). *Medición y Evaluación en la Educación y en la Psicología*. México: Compañía Editorial Continental.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and practice*, *14*(2), 5-24.

Miller, M. D. & Oshima, T.C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, *16*(4), 381-388.

Middleton, K. (2008, marzo). Assessing the Validity of a Read-Aloud Accommodation for Students with Reading-Based Learning Disabilities. Reunión Anual del National Council on Measurement in Education, Nueva York.

Montero, E. (1993). *Linguistic and cultural influences on differential item functioning for hispanic examinees in a standardized secondary level achievement test*. Tesis Doctoral no publicada, Florida State University, Miami.

Montero, E. (2001). La teoría de respuesta a los ítemes: una moderna alternativa para el análisis psicométrico de instrumentos de medición. *Revista de Matemática: Teoría y Aplicaciones. 7*(1-2), 217-228.

Muñiz, J. (1990). *Teoría de Respuesta a los Ítemes*. Madrid: Ediciones Pirámide S.A.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Ediciones Pirámide.

Nunnaly, J. & Bernstein, I. (1995). Teoría psicométrica (3a ed.). México: McGraw-Hill.

Padilla, J. L., González, A. & Pérez, C. (1998). Diferencias instruccionales y funcionamiento diferencial de los ítems: Acuerdo entre el método Mantel – Haenszel y la regresión logística. [Versión electrónica]. *Psicológica*, *19*, 201 – 215.

Pardo, C. A. (1999). *Transformaciones en las pruebas para obtener resultados diferentes*. Colombia: Servicio Nacional de Pruebas. Recuperado el 21 de octubre del 2004 de: http://200.26.128.174:8080/portalicfes/home_2/rec/arc_351.pdf

Penfield, R. D. & Camilli, G. (2006). Differential Item Functioning and Item Bias. En C.R. Rao & S. Sinharay (Eds.). *Handbook of Statistics Psychometrics*. Vol. 26 (pp.125-167). Amsterdam: Elsevier.

Schmitt, A., Holland, P. & Dorans, N. (1993). Evaluating hypotheses about differential item functioning. En P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 281–315). New Jersey: Lawrence Erlbaum Associates, Inc.

Taylor, C.S. & Lee, Y. (2008, marzo). Differential Item Functioning by Gender for Reading and Mathematics Items from Tests with Mixed Item Formats. Reunión Anual del National Council on Measurement in Education, Nueva York.

Valverde, G. (2001). La interpretación justificada y el uso apropiado de los resultados de las mediciones de logro. En P. Ravela (Ed.) Los próximos pasos: ¿Hacia dónde y cómo avanzar en la evaluación de los aprendizajes en América Latina? Recuperado el 8 de abril del 2007 de http:// www.preal.org/biblioteca.asp

Wainer, H. (1993). Model-Based Standardized Measurement of an Item's Differential Impact. En P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 123–135). New Jersey: Lawrence Erlbaum Associates, Inc.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. En P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 337 – 347). New Jersey: Lawrence Erlbaum Associates, Inc.

Manuscrito recibido en Mayo de 2007
Aceptado para publicación en Enero de 2008