Actualidades en Psicología
22, 2008, 91-113

ISSN 0258-6444

Funcionamiento diferencial del ítem en pruebas de matemática para educación media

Tania Elena Moreira Mora
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica

Dirección postal: 1437-1100 Tibás, San José, Costa Rica
Ce: tmoreira@costarricense.cr

Resumen. El tema de esta investigación concierne al funcionamiento diferencial de los ítems (FDI) y sus posibles fuentes en la población de estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad (TDAH), en la prueba de Matemática de Bachillerato en Educación Media del sistema educativo formal. En la actualidad, esta temática del FDI es fundamental en el proceso de validación de pruebas con propósitos de selección, promoción y certificación. A pesar de tal relevancia, en Costa Rica no existen estudios en este campo. Por ello, dentro de este contexto y necesidades, se pretende brindar, por un lado, un aporte metodológico al demostrar las potencialidades estadísticas del método Mantel Haenszel para identificar el FDI y, por el otro, aportar algunos presupuestos teóricos sobre FDI en la población reportada con conductas del TDAH, así como generar hipótesis para prevenir el FDI en futuras pruebas de matemática.

Palabras clave: Funcionamiento diferencial del ítem, validez, prueba de matemática, trastorno por déficit de atención con hiperactividad.

Abstract. The topic of this research concerns al differential item functioning (DIF) and its explanation for items in Math exam from the Costa Rican national exit tests in Secondary Education. Nowadays, the theme of DIF is one central issue in test validation, considering the wide use of tests for selection, promotion and certification purposes in the educational field. Despite that importance, research about identifying DIF or its sources has not been carried out in Costa Rica. In this context and considering this need, is that this study aims to contribute, in one hand, to provide a methodological approach for a more comprehensive study of DIF and its explanations, and, in the other hand, to generate some theoretical hypotheses about sources of DIF for the population of students enlisted with Attention Deficit Hyperactivity Disorder (ADHD). Hypotheses that can also be useful to control, a priori, DIF in future tests of Math.

Key Words: Functioning item differential, validity, Mathematic Test, Attention Deficit Hyperactivity Disorder

Introducción

El tema de esta investigación concierne al funcionamiento diferencial de los ítems (FDI sigla en Español y DIF en Inglés) en la prueba nacional de Bachillerato en Educación Media de matemática administrada en el año 2004 a los estudiantes egresados de la Educación Diversificada y provenientes de colegios públicos académicos diurnos del sistema educativo formal de Costa Rica.

En la actualidad este es uno de los núcleos centrales de las recientes investigaciones, debido a su amplio uso en los procesos de selección, promoción y certificación en los ámbitos educativos. A pesar de tal importancia, en Costa Rica no se han realizados investigaciones del FDI, ni de sus posibles fuentes, en el ámbito de las pruebas nacionales, ni en la población reportada con conductas del trastorno por déficit de atención con hiperactividad (TDAH). Es desde esta preocupación que se pretende brindar, por un lado, un aporte metodológico al demostrar las potencialidades del método Mantel Haenszel para identificar el FDI al proporcionar tanto un estimador de la magnitud del FDI llamado cociente de razones común MH, como una prueba de significación estadística conocida como ji-cuadrado MH con un grado de libertad; asociado a la sencillez en el cálculo e interpretación. Por el otro, aportar algunos presupuestos teóricos sobre el FDI en la población de estudiantes con conductas del TDAH, así como generar hipótesis teóricas para prevenir el FDI en futuras pruebas de matemática del programa de bachillerato.

Los primeros estudios sobre el FDI se gestaron en los años sesenta en los Estados Unidos de América a partir de los debates civiles sobre la falta de equidad en las pruebas estandarizadas, usualmente desfavorables a grupos minoritarios como: negros, hispanos y judíos. En la década de los setenta, se inicia el desarrollo de los métodos para identificar ítems potencialmente sesgados, pero es hasta mediados de los ochenta cuando se consolida un marco estadístico general que sirve de soporte para el análisis del FDI; en la década siguiente proliferan las investigaciones relacionadas al concepto de FDI y a su medición y en la actualidad es un componente clave en los estudios de validación virtualmente en todas las evaluaciones a gran escala (Penfield & Camilli, 2006). La meta específica de tales estudios fue, y continúa siendo, identificar cualquier ítem con un sesgo contra la minoría y cambiarlos o removerlos de las pruebas para crear pruebas justas y tomar decisiones importantes sobre los estudiantes (Angoff, 1993; Camilli, 1993; Haladyna, 1997; Penfield & Camilli, 2006).

En Costa Rica desde 1988 se reinstauraron las pruebas nacionales de bachillerato, establecidas inicialmente en el Código de Educación de 1943 y suspendidas en 1974. A partir de 1988 la finalidad de estas pruebas es certificar el nivel de logro académico general del currículo nacional en los estudiantes egresados de la educación diversificada, último ciclo lectivo de la educación secundaria del sistema formal de Costa Rica.

Ante el carácter decisorio de tales pruebas en la promoción de los estudiantes (equivalente a un 60% de la calificación final), así como por el peso social, académico y personal, surge la inquietud de si la interpretación de los resultados en la prueba de matemática es apropiada, o por el contrario, si existe algún impacto negativo debido a un FDI ilegítimo causado por la variancia irrelevante o la subrepresentación del constructo medido en los estudiantes. En otras palabras, esto es si la variancia en las puntuaciones de los estudiantes se debe a diferencias auténticas en el conocimiento de los examinados (impacto), o bien, a diferencias irrelevantes al constructo, que afectan la validez en la interpretación de los resultados de la prueba. De ahí que este estudio se pregunta ¿existe un funcionamiento diferencial de los ítems en la prueba ordinaria de matemática de bachillerato en educación media, aplicada a los estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad y a los estudiantes sin necesidades educativas especiales? Esto en centro educativos públicos, académicos (no técnicos) y diurnos en el 2004. Otras preguntas de investigación que se estudian aquí son las siguientes: 1) ¿Cómo identificar el funcionamiento diferencial de los ítems en la prueba ordinaria de matemática de bachillerato en educación media aplicada a los estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad y a los estudiantes sin necesidades educativas especiales de colegios públicos, académicos y diurnos en el 2004? 2) ¿Cuáles son fuentes posibles del funcionamiento diferencial de los ítems en la prueba ordinaria de matemática de bachillerato en educación media aplicadas a los estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad y a los estudiantes sin necesidades educativas especiales de colegios públicos, académicos y diurnos en el 2004? 3) ¿Cómo prevenir el funcionamiento diferencial de los ítems en futuras pruebas de matemática de bachillerato en educación media aplicada a los estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad y a los estudiantes sin necesidades educativas especiales de colegios públicos, académicos y diurnos? 4) ¿En qué medida esta investigación puede hacer aportes para el desarrollo de una teoría sobre el funcionamiento diferencial de los ítems en la prueba de matemática de bachillerato en educación media aplicada a los estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad?

El objetivo principal de esta investigación consiste en evaluar el funcionamiento diferencial de los ítems en la prueba de matemática de

bachillerato en educación media aplicada en el año 2004 a los estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad y a los estudiantes sin necesidades educativas especiales provenientes de colegios públicos, académicos y diurnos. De manera específica nos proponemos, en primer lugar, identificar el funcionamiento diferencial de los ítems en la prueba ordinaria de matemática de bachillerato en educación media mediante la aplicación de los métodos empíricos la diferencia p estandarizada y el Mantel-Haenszel. En segundo lugar, se busca determinar las posibles fuentes del funcionamiento diferencial de los ítems en la prueba ordinaria de matemática de bachillerato en educación media aplicada a los estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad y a los estudiantes sin necesidades educativas especiales de colegios públicos, académicos y diurnos en el 2004. También interesa generar hipótesis teóricas para prevenir el funcionamiento diferencial de los ítems en futuras pruebas de matemática de bachillerato en educación media. Y finalmente, proponemos algunas ideas para el desarrollo de una teoría sobre el funcionamiento diferencial de los ítems en la prueba de matemática de bachillerato en educación media en la población de estudiantes reportados con conductas del trastorno por déficit de atención con hiperactividad.

En el contexto educativo de Costa Rica se entiende por educación especial el "conjunto de apoyos y servicios a disposición de los alumnos con necesidades educativas especiales, ya sea que los requieran temporal o permanentemente" (MEP, 1998, p. 5). Este es un asunto de interés en esta investigación, debido a que la población analizada está conformada por dos subgrupos: el de referencia (GR) constituido por los estudiantes sin necesidades educativas especiales y el focal (GF) conformado por los estudiantes reportados con conductas del TDAH. Conforme con la ley 7600 sobre la Igualdad de Oportunidades para las Personas con Discapacidad, vigente desde el 29 de mayo de 1996, los estudiantes con necesidades educativas especiales se integran al sistema regular con los servicios de apoyo requeridos y las adecuaciones curriculares necesarias para garantizar una igualdad de oportunidades educativas y una educación de igual calidad.

De acuerdo con las políticas de acceso, una adecuación curricular es la acomodación o ajuste de la oferta educativa a las características y necesidades de cada alumno, con el fin de atender las diferencias individuales. Estas acomodaciones pueden ser de acceso, no significativas y significativas. Los estudiantes del grupo focal constituían el 82.5% de los estudiantes con adecuación curricular en ese año y recibieron adecuaciones no significativas, en particula, una hora adicional para resolver la prueba y la presencia de un tutor especialista en la administración de la prueba.

En general, con tales adecuaciones se pretende satisfacer las necesidades educativas especiales de los estudiantes reportados con

conductas del TDAH, puesto que por su déficit presentan un mayor grado de dificultad para aprender con respecto al promedio de sus compañeros de la misma edad. Este grupo se caracteriza, usualmente, por la dificultad para mantener la atención en sus actividades cotidianas y académicas, según Villalobos y Morales (2002), en tres áreas específicas: capacidad de concentración, capacidad de control de impulsos y, en algunos casos, en el nivel de actividad.

Este es uno de los trastornos que incluye en las disfunciones del desarrollo neurobiológico infantil. En términos psicológicos, el TDAH comprende "a un grupo heterogéneo de manifestaciones clínicas cuyas materializaciones conductuales más visibles, a grandes rasgos, son la hiperactividad, la impulsividad y la dificultad para mantener la atención" (Moreno, 2001, p. 81). De acuerdo con esta definición, el TDAH involucra un conjunto de características conductuales visibles, que tienden a la desadaptabilidad de niños, niñas y jóvenes tanto en su ambiente académico como familiar.

En cuanto a las características de la prueba de matemática aplicada tanto a los estudiantes del GR como al GF, en primera instancia, se enuncia que corresponde al modelo con referencia a normas, asentado en el principio de la curva normal y en la discriminación entre los estudiantes en términos del nivel relativo de aprendizaje en diferentes áreas temáticas; por tanto, sus resultados se pueden utilizar para certificar. Las normas son datos estadísticos que relacionan la calificación de una persona con las puntuaciones de otros. Por lo tanto, una información normativa no nos indica en forma automática si la calificación de un sujeto está por arriba o por debajo del nivel en que debiera estar; nos señala únicamente su ejecución en comparación con otros (Nunnally & Bernstein, 1995; Mehrens & Lehmann, 1982). Esta comparación de las puntuaciones obtenidas por los examinados depende de las propiedades psicométricas de la prueba de matemática, particularmente de la confiabilidad, la unidimensionalidad y la validez.

La confiabilidad o fiabilidad se evidencia en la consistencia interna y en la estabilidad temporal de las puntuaciones. La primera definición recoge el grado de coincidencia existente entre los elementos que la componen y la estabilidad en el tiempo alude a la capacidad del instrumento para arrojar las mismas mediciones cuando se aplica más de una vez a los mismos sujetos (Anastasi & Urbina, 1998; Pardo & Ruiz, 2002). En esta investigación se utilizó el modelo de Alfa de Cronbach para determinar la consistencia interna de las puntuaciones de la prueba de matemática.

También las pruebas nacionales de bachillerato en educación media se basan en el supuesto de la unidimensionalidad, al precisar la medición de un solo constructo, variable latente o factor. En el caso de la prueba de matemática con la medición de los contenidos y objetivos instruccionales del

temario unificado del 2004. La técnica más utilizada para comprobar la unidimensionalidad ha sido el análisis de la estructura factorial por ser un procedimiento estadístico que determina las interrelaciones de los datos conductuales y reduce el número de variables o categorías en factores, en cuyos términos puede describirse el desempeño de cada individuo, a un número relativamente pequeño de factores (llamados compuestos, constructos, dimensiones, índices, ejes o rasgos homogéneos) que se distinguen por ser independientes, por los que se trata de una técnica de reducción de la dimensionalidad de los datos y determinar así, el grado en que las medidas hipotetizadas de un constructo miden lo mismo (Anastasi & Urbina, 1998; Martínez, 2005; Nunnallly & Bernstein, 1995; Pardo & Ruiz, 2002).

Finalmente, los ítems de las pruebas nacionales de bachillerato deben ser sometidos a un proceso de validación. Messick (1995, p.5) entiende por validez "…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". Conforme con los estándares para las pruebas educativas y psicológicas establecidos en 1999 por la "American Educational Research Association" (AER), la "American Psycological Association" (APA) y el "National Council on Measurement in Education" (NCME), el proceso de validación involucra la acumulación de evidencia que proporciona una base científica para la interpretación y la relevancia de las puntuaciones de la prueba.

La esencia de esta perspectiva conceptual es la integración de los tres tipos de evidencias tradicionales: contenido, criterio y constructo por el hilo unificador de la validez de constructo, para apoyar lo adecuado de las interpretaciones, usos y consecuencias sociales de las puntuaciones de una prueba. Sin embargo, puede caerse en la tentación de confiar en solo una categoría de evidencia, por ello Messick (1989) plantea dos facetas interrelacionadas: una es la fuente de la justificación de la prueba, basada en el estudio de la evidencia que apoya el significado de la nota o en las consecuencias que contribuyen a la valoración de la nota. La otra faceta es la función o el resultado de la prueba, siendo la interpretación o el uso. Dentro de esta perspectiva unificada, el análisis del FDI es elemental por aportar evidencia empírica del grado en que las puntuaciones miden apropiadamente un constructo.

El funcionamiento diferencial del ítem, de manera frecuente, se ha confundido con sesgo, especialmente por el doble sentido de sesgo: uno social y otro estadístico. Al respecto Angoff (1993) señala que la consecuencia ha sido una confusión innecesaria, por lo que algunas sugerencias apuntan al uso de un término para describir el juicio y evaluación de sesgo con un sentido social y otro término para referirse a las

observaciones estadísticas. Precisamente la expresión FDI se utiliza para las propiedades estadísticas en diferentes grupos.

En el ámbito de la teoría clásica de los test (TCT) se ha utilizado el término de sesgo para rotular los ítems que tienen índices de dificultad o discriminación diferentes en los grupos comparados (Andriola, 2001). Mientras que, desde un punto de vista psicométrico, un ítem presenta un 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; Attorresi, Galibert, Zanelli, Lozzia & Aguerri, 2003; Camilli, 1993; Hidalgo, Galindo, Inglés, Campoy & Ortiz, 1999; Muñiz, 1990; Padilla, González & Pérez, 1998; Penfield & Camilli, 2006). 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 y Thayer, 1993; Hidalgo et al., 1999; Montero, 1993). En suma, un ítem presenta un FDI si en igualdad de condiciones los examinados pertenecientes al grupo de referencia sistemáticamente tienen una mayor probabilidad de responder correctamente el ítem que los estudiantes del grupo focal.

Es en la década de los setenta cuando surgen múltiples procedimientos estadísticos rigurosos para la detección del FDI, clasificados en dos grandes categorías. Uno comprende los métodos empíricos (también conocidos como métodos de invariancia condicional observada o métodos condicionales) basados en las puntuaciones observadas en la prueba desde la perspectiva de la teoría clásica. El otro incluye los métodos teóricos (también conocidos como métodos de invariancia condicional no observada o métodos incondicionales) fundamentados en modelos matemáticos como los de la teoría de respuesta a los ítems (TRI) por utilizar las estimaciones de la habilidad (θ), según el modelo más adecuado a los datos (Andriola, 2002, 2003; Hidalgo, López & Sánchez, 1997; Montero, 1993; Wainer, 1993). En esta investigación se escogieron el Mantel-Haenszel y la diferencia p estandarizada, ambos métodos empíricos, para identificar los ítems con un FDI; puesto que la prueba de matemática de bachillerato en educación media es construida bajo un modelo clásico y conforme con Hidalgo et al. (1997) la selección del método depende de las características del modelo de medida, de tal forma que si la prueba ha sido construida bajo el modelo clásico se pueden utilizar los métodos empíricos.

El método Mantel-Haenszel proporciona tanto un estimador de la magnitud del FDI llamado cociente de razones común MH (α_{MH}) como una prueba de significación estadística conocida como ji-cuadrado MH (χ^2_{MH}) con un grado de libertad. Este método se basa en la comparación de las frecuencias observadas y esperadas de aciertos y errores en un ítem por

sujetos que perteneciendo a distintas poblaciones (grupo focal y de referencia) muestran el mismo nivel de puntuación en la prueba (Andriola, 2002, 2003; Elosúa & López, 1999; Hidalgo et al., 1999; Longford, Holland & Thayer, 1993). En el cálculo de este estadístico la variable latente θ se divide en K intervalos de habilidad y se construyen K tablas de contingencia 2 x 2 para cada ítem. En cada una de las tablas los sujetos son clasificados según el grupo de pertenencia (focal o referencia) y las posibles respuestas al ítem (Andriola, 2002, 2003; Hidalgo et al., 1997, 1999; Longford, Holland & Thayer, 1993).

El cociente de razones (α_{MH}), también conocido como razón de productos cruzados ("odds ratio"), expresa el cociente o razón entre la probabilidad de acertar el ítem contra la probabilidad de fallarlo en el grupo focal y la probabilidad de responderlo correctamente contra la probabilidad de fallarlo en el grupo de referencia (Andriola, 2002; Hidalgo et al., 1997, 1999). El cociente de razones Mantel y Haenszel se obtiene de la siguiente expresión:

$$\alpha = \frac{A_k D_k / T_k}{C_k B_k / T_k} \qquad (1)$$

El valor del α_{MH} puede oscilar entre 0 y ∞. Si el α_{MH} es mayor que 1 favorece al grupo de referencia y si es menor que 1, indica que el grupo focal muestra un desempeño más alto en comparación con el de referencia (Andriola, 2002; Bishop, Sharairi, Swift, Wa Lei, & Domaleski, 2006). Sin embargo, por cuestiones prácticas, Penfield (2006) propone en su programa DIFAS, utilizado en esta investigación, una transformación de los "odds ratio" (α_{MH}) con una distribución normal asintótica donde muestran un comportamiento logarítmico. Si los "odds ratio" es mayor que 1 entonces el logaritmo "odds ratio" es positivo lo que indica un FDI a favor del grupo de referencia y si los 0 < "odds ratio" < 1 entonces el logaritmo "odd ratio" es negativo lo que señala un FDI a favor del grupo focal. En el caso de los ítems dicotómicos, DIFAS también proporciona una clasificación de los coeficientes basados en el esquema del Educational Testing Service (Carvajal & Poggio, 2006).

El "Educational Testing Service" (ETS) propuso una escala jerárquica para los distintos valores del coeficiente ΔMH (logaritmo de los "odds ratio" en una métrica delta) o también conocido como "Mantel Haenszel delta difference" (MH D-DIF), de acuerdo con su magnitud. Esta clasificación está basada en dos factores, el valor absoluto del MH D-DIF y si este valor es significativo a un nivel de probabilidad del .05 (p =.05); ambos son importantes, aún en casos de valores muy pequeños de FDI pero

estadísticamente significativos, porque el análisis se ha basado en un gran número de examinados (Zieky, 1993). Las tres categorías de MH D-DIF se han etiquetado con las letras A, B y C (Andriola, 2002; Dorans & Holland, 1993; Longford et al., 1993; Ziecky, 1993).

- Categoría A: Ítems con valores MH D-DIF no significativamente diferentes de 0 ($p = 0.05$) o con valores absolutos menores que 1 (unidad delta) serán considerados ítems con un FDI despreciable o insignificante;
- Categoría B: Ítems con valores MH D-DIF significativamente diferentes de 0 ($p = 0.05$) y con valores absolutos iguales o mayores que 1 pero no significativamente mayores que 1 (opción 1); o con valores absolutos iguales o mayores que 1 pero menores a 1.5 (opción 2) serán considerados ítems con FDI moderado;
- Categoría C: Ítems con valores MH D-DIF significativamente mayores que 1 ($p = 0.05$) y con valores absolutos iguales o mayores que 1.5 serán considerados ítems con FDI severo.

Otro modelo empírico ampliamente utilizado para la identificación del FDI, especialmente por el ETS ha sido la diferencia p estandarizada (STD P-DIF, por sus siglas en inglés) o método estandarizado. De acuerdo con dicho método, un ítem tendrá un FDI cuando el rendimiento esperado para individuos con igual grado de habilidad, pero provenientes de distintos grupos es diferente (Dorans & Holland, 1993; Montero 1993). Con este método se calcula el índice de discrepancia entre los grupos con respecto al rendimiento en un ítem ("p difference") con base en la siguiente expresión matemática descrita por Montero (1993).

$$\Sigma [K_s (p_{fs} - p_{bs})] / \Sigma K_s \qquad (2)$$

Donde:

p_{fs} = la proporción de respuestas correctas en el grupo focal (minoría) en el nivel de habilidad "s".

p_{bs} = la proporción de respuestas correctas del grupo base (mayoría) en el nivel de habilidad "s".

K_s = es el factor de peso para cada nivel de puntuación "s", es decir, es el número de personas del grupo focal en el nivel "s".

La STD P-DIF es un índice que puede tomar un valor entre -1 y 1 (ó -100 y 100), cuya dirección es dada por el signo. Los valores positivos señalan que el ítem favorece al grupo focal; mientras que los negativos indican un FDI contra el grupo focal. Los valores de la STD P-DIF se organizan en la siguiente escala jerárquica, propuesta por el "Educational Testing Service",

de acuerdo con su magnitud (Andriola, 2003; Dorans & Holland, 1993; Montero, 1993):

- Categoría A: los ítems cuyos índices oscilan entre -0.05 y 0.05 tienen un FDI considerado insignificante.
- Categoría B: los ítems cuyos índices oscilan entre -0.10 y -0.06 y entre 0.06 y 0,10 tienen un FDI moderado y deben ser revisados para asegurarse que no es efecto de un descuido.
- Categoría C: los ítems con valores fuera del rango [-0.10, 0.10] son inusuales y deben ser examinados cuidadosamente al presentar un FDI severo.

En definitiva, el uso de ambos métodos empíricos en esta investigación responde, por un lado, a la tendencia de utilizar dos o más procedimientos y, por el otro, a identificar el método más eficaz y robusto para detectar el FDI.

Método

Esta investigación es no experimental puesto que la investigadora se limitó a la observación del FDI en la prueba de matemática, sin introducir ninguna alteración en el tratamiento educativo, en la administración y confección de la prueba, ni en las puntuaciones obtenidas por los estudiantes. Además, por enfocarse en un problema de investigación que nunca se ha abordado en nuestro contexto educativo e investigativo se clasifica como exploratoria y por el marco temporal califica como un diseño transversal debido a que los datos fueron obtenidos de la prueba aplicada en la única convocatoria ordinaria del 2004.

Participantes

El estudio se llevó a cabo con la población de estudiantes que realizó la prueba de matemática en la convocatoria ordinaria del 2004 provenientes de 217 colegios públicos académicos diurnos del sistema educativo formal de Costa Rica, específicamente entre dos grupos que realizaron la prueba ordinaria: el de referencia constituido por 14132 estudiantes que no tenían necesidades educativas especiales y el focal conformado por 493 alumnos reportados con conductas del TDAH, quienes realizaron las pruebas con adecuaciones curriculares.

Procedimientos

El estudio se realizó en dos etapas principales. La primera caracterizada por su énfasis exploratorio, donde se utilizaron dos técnicas cualitativas: la observación participante en un grupo de undécimo año de un colegio público académico diurno para observar a los estudiantes reportados con el TDAH y el grupo de discusión con los estudiantes observados. La aplicación de ambas técnicas respondía al interés por comprender las manifestaciones del TDAH en las vivencias cotidianas de estos jóvenes, cuyos hallazgos contribuyeron a la formulación de algunas hipótesis teóricas. Esto porque en los estudios referentes al FDI se pueden enunciar dos tipos de hipótesis: las teóricas y las empíricas. En el caso concreto de esta exploración, se optó por las teóricas sustentadas tanto en el análisis de los datos recolectados con estas dos técnicas como en el conocimiento teórico del TDAH y en los descubrimientos de otras investigaciones.

Posteriormente, en esta primera etapa se calculó el coeficiente Alfa de Conbach para determinar el grado de consistencia interna de los resultados. Además el análisis de la estructura factorial para comprobar el supuesto de unidimensionalidad y, finalmente, la detección empírica del FDI utilizando el procedimiento de la diferencia p estandarizada y el Mantel Haenszel.

La segunda etapa tiene un carácter interpretativo y se orientó a la indagación de las posibles fuentes del FDI, mediante la técnica de jueces constituidos por especialistas en matemática y educación especial para explicar los posibles atributos irrelevantes presentes en los ítems de selección única detectados estadísticamente con un FDI alto en ambas pruebas. En general, el procedimiento de análisis se resume en los siguientes pasos:

1. Aplicación de la técnica de la observación participante en un grupo de estudiantes de undécimo año de un colegio público diurno académico.
2. Utilización de la técnica grupo de discusión con estudiantes reportados con conductas del TDAH.
3. Elaboración en el programa del SPSS, versión 12.0 para Windows, de las bases de datos de la población que realizó la prueba de matemáticas para realizar el análisis de la confiabilidad y la estructura factorial.
4. Análisis de la confiabilidad con el programa del SPSS.
5. Análisis de la estructura factorial con el programa del SPSS.
6. Aplicación del método de la diferencia p estandarizada con el programa Microsoft EXCEL y el SPSS.
7. Análisis de los datos con el método Mantel Haenszel usando el programa informático "Differential Item Functioning Analysis System" (DIFAS).
8. Aplicación de la técnica de los jueces para analizar las posibles fuentes del FDI.

En suma con esta propuesta metodológica, fundamentada en el método hipotético deductivo, se analizó el FDI en las pruebas objeto de estudio con base en la triangulación de la teoría sobre el TDAH, las evidencias estadísticas y los datos cualitativos.

Resultados

En primer lugar se identificó el FDI en los ítems de la prueba de matemática; no obstante, antes de aplicar los dos métodos empíricos, fue necesario analizar previamente la propiedad de la confiabilidad y el supuesto de unidimensionalidad.

La consistencia interna inicial de la prueba de matemática, construida con 60 ítems de selección única y calculada con la totalidad de examinados de ambos grupos, fue de 0.843. En este análisis se eliminaron cuatro ítems por dos razones. Una por afectar la precisión de los resultados en términos del Alfa de Cronbach y la otra para evitar inconsistencias en la detección del FDI. El valor final alcanzado con los 56 ítems fue de 0.85. De acuerdo con Nunnally y Bernstein (1995), el nivel satisfactorio de confiabilidad depende de cómo se use una medida, de tal manera que en las primeras etapas de una investigación de validación predictiva o de constructo, un índice de 0.70 es aceptable.

El análisis de la estructura factorial fue de carácter exploratorio, en el cual se asumió que los factores son conceptualmente independientes pero correlacionados, por ello se recurrió a una rotación oblicua en el método de extracción de componentes principales. Los indicadores utilizados en estudios del FDI para medir la unidimensionalidad han sido el porcentaje de variancia explicada por el factor común y el gráfico de sedimentación.

El porcentaje de variancia explicada por el primer componente en el grupo focal fue de 12.59 y en el grupo de referencia, de 11.26. Ambos valores aceptables para considerar cierto grado de unidimensionalidad, pues como lo destaca Andriola (2002) es una cuestión de grado o sea cuánta más variancia explique el primer factor, más unidimensionalidad existirá. Por otra parte, de acuerdo con los gráficos de sedimentación se observa el predominio del primer componente, aunque a partir del punto de transición donde la curva cambia de un descenso fuerte a uno más gradual se retienen dos componentes en el GR, cuya correlación fue de 0.451. Mientras que en el GF la correlación entre el primer componente con el segundo fue de 0.405 y con el tercero la asociación fue menor, de 0.325.

En cuanto a la detección de los ítems con un FDI alto se destaca que el método Mantel Haenszel fue más sensible y preciso por incorporar la prueba estadística χ^2_{HM}, cuyos resultados coincidieron parcialmente con los

índices STD P DIF. A partir de este análisis empírico con los 56 ítems, se detectó un 17.9% (10 ítems) con un FDI alto, de los cuales 5.4% resultó favorable al grupo focal (3 ítems) como se muestra en la siguiente tabla.

Tabla 1
Resumen del análisis del funcionamiento diferencial de los ítems en la prueba de matemática

Ítem	Mantel Haenszel			Diferencia p estandarizada	
	χ^2_{MH}	Log odds ratio	ETS	Índice de discrepancia	ETS
9	4.5238	0.2149	A		
10	3.8545	0.1939	A		
16	8.7844	-0.2943	A	0.064	B
18	5.1199	-0.4306	A		
20	6.7964	0.3339	A		
26	6.7687	0.3052	A	-0.052	A
34	4.3479	0.267	A		
37	4.4699	-0.2129	A		
41	8.8674	0.3887	A	-0.055	A
45	8.035	0.3561	A	-0.054	A

Según la escala del "Educational Testing Service" (ETS), 9 ítems se clasificaron en la categoría A (irrelevante) y solo uno con una magnitud moderada (B). Sin embargo, esta clasificación no le resta importancia a la significancia estadística, puesto que todos resultaron con un FDI alto.

El segundo objetivo de esta investigación compete a la determinación de las posibles fuentes del comportamiento diferencial de los diez ítems detectados en el análisis empírico. La tendencia general en este tipo de estudios ha sido la combinación de los procedimientos estadísticos con el criterio de jueces para identificar las fuentes de variancia irrelevante del constructo.

Los ítems fueron analizados por tres especialistas en el área de la matemática y dos de la educación especial, quienes por su experiencia y conocimientos, argumentaron que los siete ítems con un FDI alto en contra del grupo focal, se caracterizaban por: el empleo de vocabulario impreciso, redacción poco clara, la transición del lenguaje verbal al algebraico, procedimientos complejos por la cantidad de cálculos, conceptos y detalles, dibujo de figuras, colocación espacial de las expresiones y figuras matemáticas. Con respecto a los tres ítems con un FDI alto a favor del grupo focal, los jueces apuntaron que probablemente: la estructura sencilla del ítem, el empleo de vocabulario exacto, la medición de un solo concepto y la corta extensión fueron las características que favorecieron a este grupo.

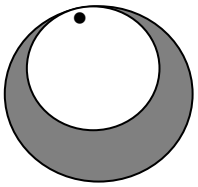
El tercer objetivo correspondiente al análisis de las cuatro hipótesis teóricas se basó en la evidencia empírica de los diez ítems identificados con un FDI alto, según los resultados del método Mantel Haenszel y de la

diferencia p estandarizada, así como del análisis de la estructura factorial para examinar la posibilidad de multidimensionalidad del ítem.

En este sentido sometimos a prueba nuestra primera hipótesis, según la cual los ítems de matemática que incluyen figuras geométricas y gráficas que adicionan múltiples datos en el planteamiento del encabezado favorecen un funcionamiento diferencial alto en contra de los estudiantes reportados con conductas del TDAH.

En cuanto a la evidencia empírica sólo se detectó un ítem con un FDI alto en contra del grupo focal que incluía una figura geométrica en el encabezado, como se ilustra en el ejemplo 1. No obstante, fue un criterio compartido por los cinco jueces de matemática que las figuras ayudaron a ubicar a los examinados en el planteamiento de la pregunta, bajo la condición que ilustre de manera precisa el problema, sin detalles innecesarios que los distraigan o confundan. En el caso de este ejemplo 1 posiblemente, según los expertos, desfavoreció a los estudiantes reportados con el TDAH por la redacción poco clara del encabezado y no particularmente por la figura, calificada de específica y sencilla.

Ejemplo 1.
Considere la siguiente figura.



Sea R la longitud del radio de una de las circunferencias. Si la longitud del radio de la otra circunferencia es $\frac{3}{4}R$, entonces el área de la región destacada con gris equivale a

A) $R\pi$

B) $\frac{R\pi}{2}$

C) $\frac{R^2\pi}{16}$

D) $\frac{7R^2\pi}{16}$

Esta posición se reafirma con otro ítem que resultó con un FDI alto a favor del grupo focal, cuya gráfica fue considerada específica, igual que la figura del ejemplo 1, lo que favoreció a los estudiantes a situarse directamente en el problema. Asimismo, cabe agregar que dos ítems resultaron con un FDI alto en contra del grupo focal, debido, entre otras posibles fuentes, a que el estudiante debía dibujar la gráfica y la figura geométrica.

En suma, en esta primera hipótesis la evidencia empírica apunta más bien a la importancia del apoyo visual de las figuras geométricas y las gráficas para los estudiantes, especialmente para los que presentan conductas del TDAH, bajo la cláusula que ilustren de manera puntual; sin adicionar datos innecesarios ni sobrecargar con otras imágenes como recuadros con información, que lo harían visualmente complicado para los examinados del grupo focal.

Nuestra segunda hipótesis plantea que los ítems de matemática que combinan múltiples datos y conceptos en el encabezado para que el estudiante realice los procedimientos y encuentre la respuesta correcta contribuyen a un funcionamiento diferencial alto en contra de los estudiantes reportados con conductas del TDAH.

En el análisis del FDI se obtuvo evidencia empírica para apoyar esta hipótesis, puesto que de los siete ítems con un FDI en contra del grupo focal, tres se caracterizaban por incorporar múltiples datos y conceptos en el planteamiento del ítem tales como: i) datos para calcular el largo del rectángulo y resolver la ecuación, ii) múltiples conceptos (concavidad, vértice, intersección, función decreciente y función cuadrática) para dibujar la gráfica y discriminar entre dos proposiciones y iii) conceptos de plano, distancia, circunferencia, secante, concéntricas y radio para identificar la circunferencia de la figura geométrica, como se muestra en el ejemplo 2.

Ejemplo 2

En un mismo plano, la distancia entre los centros de dos circunferencias es 10. Si "R" representa la medida del radio de una de ellas y "10 – R" representa la medida del radio de la otra, entonces se cumple que las circunferencias son

A) secantes.

B) concéntricas.

C) tangentes interiormente.

D) tangentes exteriormente.

Es importante subrayar que la combinación de múltiples datos y conceptos se identificó como una posible fuente, mas los expertos igualmente reconocieron otras probables fuentes del FDI en un mismo ítem, que en conjunción con lo propuesto en esta hipótesis, contribuyeron a un comportamiento diferencial en contra de los estudiantes reportados con conductas del TDAH.

Según nuestra hipotésis 3 los ítems de matemática que incluyen en el encabezado varias proposiciones para clasificarlas como falsas o verdaderas favorecen un funcionamiento diferencial alto en contra de los estudiantes reportados con conductas del TDAH.

Únicamente un ítem resultó con un comportamiento diferencial alto en contra del grupo focal, como se ilustra en el ejemplo 3. No obstante, de acuerdo con los expertos la estructura del ítem resulta más complicada cuando el estudiante debe discriminar entre dos proposiciones, ya que implica la realización de varios cálculos para reconocer cuál es la verdadera.

Ejemplo 3

Sea (2.9) el vértice y (0.3) el punto de intersección con el eje "y" de la gráfica de una función cuadrática f; considere las siguientes afirmaciones.

I. f es cóncava hacia abajo.

II. f es estrictamente decreciente en $]\,2\,, +\infty\,[$.

De ellas, ¿cuáles son **verdaderas**?

A) Ambas.

B) Ninguna.

C) Solo la I.

D) Solo la II.

En concreto, la inclusión de dos proposiciones para clasificarlas como verdaderas es una característica que aumenta la complejidad del ítem, lo que desfavorece a los estudiantes reportados con conductas del TDAH, por su deficiencia para abstraer, discriminar y organizar tanta información. Sin embargo, en la prueba de matemática había un total de ocho ítems con esta característica y únicamente se detectó un ítem con FDI alto, por ende, la evidencia no es suficiente para sostener esta presunción.

Con la hipótesis 4 planteamos que si existe una configuración factorial diferente en la prueba de matemática aplicada a los estudiantes reportados

con conductas del TDAH y a los estudiantes sin necesidades educativas especiales, esa diferencia es una fuente de FDI.

Con respecto al análisis de la estructura factorial se aclara que el principal objetivo fue obtener evidencia de la unidimensionalidad. Adicionalmente, con este análisis se puede identificar aquellos ítems que miden un componente o factor secundario en alguno de los grupos comparados, tal diferencia factorial sería una posible fuente del FDI. Específicamente, si la carga factorial del ítem es inferior a 0.30 se evidencia que el ítem no representa el constructo principal de cada prueba, ya sea en uno o en los dos grupos de examinados.

A partir de este análisis, se obtuvo evidencia de multidimensionalidad como fuente del FDI en dos ítems. En el primero el ítem presentó un FDI a favor del grupo focal y la carga factorial fue alta (0.326) únicamente en este grupo; probablemente, en el grupo de referencia no correlacionó alto con el componente principal por medir otro secundario.

El segundo ítem resultó con un FDI a favor del grupo de referencia y presentó una carga factorial de 0.346 en el componente, mientras que en el focal fue menor a 0.30, probablemente por medir algún componente o factor secundario, tal vez, la habilidad de transformar el lenguaje verbal al matemático. En suma, esta evidencia empírica permite aceptar la hipótesis de multidimensionalidad, al encontrarse diferencias en la estructura factorial entre el grupo focal y el de referencia.

Pensando en el desarrollo de una teoría sobre el FDI en la población del grupo focal, esta contribución se cimienta en los planteamientos teóricos sobre el TDAH, en los hallazgos cualitativos y, concluyentemente, en la evidencia estadística que permitió generar algunas hipótesis empíricas que se deben comprobar en futuros estudios de tipo confirmatorio y experimental. Específicamente, con los ítems de matemática detectados con un funcionamiento diferencial ilegítimo desfavorable a los examinados reportados con conductas del TDAH se descubrieron como probables fuentes de FDI:

1) La conversión del lenguaje verbal al lenguaje algebraico.
2) Procedimientos de resolución complejos debido a la combinación de múltiples datos, conceptos y cálculos matemáticos, sumado al reconocimiento de proposiciones verdaderas.
3) La elaboración de gráficas o figuras geométricas en el procedimiento para encontrar la respuesta correcta.
4) El uso de términos imprecisos en el encabezado para indicar la ejecución de alguna operación o relacionar expresiones matemáticas.
5) La redacción confusa del planteamiento del problema en el encabezado.

6) La saturación visual del encabezado debido a la colocación espacial de los datos, el texto y las expresiones matemáticas.
7) La elaboración de distractores visualmente muy semejantes.

En general, los examinados del grupo focal tienen un nivel de abstracción y un funcionamiento cerebral diferente al de referencia, así como mayores impedimentos para organizar el razonamiento, los procedimientos y cálculos matemáticos cuando: deben resolver un ítem complejo por la cantidad de información y cálculos, la redacción del encabezado y opciones es confusa, o bien, cuando el ítem visualmente está saturado.

Conclusiones

En primera instancia, se resaltan dos premisas asumidas en esta investigación. La primera se fundamenta en las evidencias empíricas obtenidas en múltiples estudios realizados en Estados Unidos, que han comprobado el aporte de las adecuaciones en el desempeño de los estudiantes en las pruebas a gran escala, lo que en definitiva sustenta la postura asumida en esta investigación que el propósito de las adecuaciones es medir con una mayor exactitud los conocimientos matemáticos de los estudiantes del grupo focal, comparables con las obtenidas en el grupo de referencia, al eliminar el efecto de su trastorno o discapacidad en la ejecución de la prueba. Entonces se admite la premisa según la cual las adecuaciones curriculares de acceso y no significativas incrementan la validez de las interpretaciones de los resultados obtenidos por los estudiantes del grupo focal, al eliminar los impedimentos irrelevantes en la medición del constructo en la prueba de matemática. La segunda premisa es que los resultados de cualquier investigación en esta área dependen del diseño de la investigación, del tamaño de las muestras, del tipo de discapacidad y de los modelos estadísticos utilizados, por tanto, los hallazgos encontrados en este estudio se deben contextualizar e interpretar dentro del modelo teórico y metodológico propuesto en este estudio.

En cuanto al primer problema de investigación concerniente a cómo identificar el funcionamiento diferencial del ítem se concluye que es indispensable realizar previamente el análisis de la confiabilidad y de la estructura factorial para cumplir con el estándar de la consistencia interna de los resultados y con el supuesto a la unidimensionalidad antes de aplicar los métodos de detección del FDI.

El análisis de la estructura factorial evidenció indicios de cierta desigualdad en la medición del mismo constructo entre el grupo focal y el de referencia. De ahí se concluye que estas variaciones en la configuración

factorial y en los agrupamientos de los ítems en los componentes principales, demostraron diferencias psicométricas que se deben de comprobar con un análisis de tipo confirmatorio. No obstante, se comprobó el supuesto de unidimensionalidad sustentable en varios indicadores como el porcentaje de variancia explicada por el primer componente y el grado de correlación entre los componentes.

Otra conclusión es la necesidad de comprobar el grado de coincidencia en la identificación de los ítems con un funcionamiento diferencial con el empleo de diferentes métodos empíricos y teóricos. Concretamente, en los resultados del Mantel Haenszel y la diferencia p estandarizada no hubo una total convergencia; hallazgo esperable debido a la falta de consistencia entre los métodos empíricos. Entonces, ante el problema de cómo identificar el FDI se concluye, en primera instancia, que la aplicación de los métodos empíricos depende del cumplimiento del estándar de confiabilidad y del supuesto de unidimensionalidad y segundo de la factibilidad y manejo de paquetes estadísticos, comprobándose las potencialidades del Mantel Haenszel sobre la diferencia p estandarizada.

El segundo problema de esta investigación se enfocó a determinar las posibles fuentes del funcionamiento diferencial de los ítems detectados en el análisis empírico. En esencia, la respuesta a esta interrogante se logró con el aporte de un equipo de jueces, quienes examinaron el contenido y el formato de los ítems para intentar dar una explicación de este fenómeno. Particularmente, los jueces señalaron que las potenciales fuentes del comportamiento diferencial alto en contra de los examinados del grupo focal se encontrarían en características como: el empleo de vocabulario impreciso, la redacción poco clara, la transición del lenguaje verbal al algebraico, los procedimientos complejos por la saturación de cálculos y conceptos, la colocación espacial de las expresiones y figuras matemáticas y, finalmente, el trazo de figuras geométricas o gráficas. En conclusión, tales atributos no son relevantes en la medición de los conocimientos matemáticos, pues el estudiante reportado con conductas del TDAH puede dominar los conceptos y los procesos, pero cuando se enfrenta ante un ítem confuso e impreciso por su redacción, o bien, saturado de información no logra focalizar su atención en el problema y organizar el procedimiento de resolución; lo que afecta, sin duda alguna, sus probabilidades de responder correctamente, a pesar de ser igualmente competente que los examinados del grupo de referencia.

En definitiva, a partir de los aportes de los jueces se encontró la respuesta a este segundo problema. No obstante, estas explicaciones deben asumirse en términos de posibilidades, por la naturaleza exploratoria de esta investigación, debido a la carencia de estudios y de una teoría acerca de este fenómeno. Precisamente en esto radica la principal razón teórica de esta

investigación, descubrir las fuentes del FDI en este grupo minoritario con la finalidad de lograr una equidad psicométrica en la prueba de matemática.

El tercer problema concierne, justamente, a cómo prevenir el FDI en futuras pruebas de matemática, lo que implicó establecer algunas hipótesis teóricas, enfocadas principalmente en aspectos de estructura y redacción del ítem como posibles atributos irrelevantes que contribuyeron a un FDI alto en contra del grupo de examinados del grupo focal, pero no se exploró el contenido de los ítems, probable fuente como se ha evidenciado en algunas de las investigaciones revisadas.

A partir del análisis de estas hipótesis, basado tanto en la evidencia estadística de los siete ítems identificados con un FDI alto en contra del grupo focal como en los criterios externados por los jueces, se concluye que fue la conjugación de varios atributos irrelevantes en un ítem las probables fuentes de un FDI alto en contra del grupo focal. Entonces, ante la interrogante de cómo prevenir el FDI alto en contra del grupo focal, se obtuvieron evidencias estadísticas de ciertos atributos, específicamente los planteados en la segunda hipótesis, que se deben evitar en la construcción de ítems en futuras pruebas de matemática.

En el último problema de esta investigación se planteó en qué medida este estudio puede hacer aportes para el desarrollo de una teoría sobre el funcionamiento diferencial de los ítems de matemática aplicada al grupo poblacional reportado con conductas del TDAH. Indiscutiblemente, la respuesta a esta interrogante se limita al marco referencial de esta investigación y se relaciona con los planteamientos teóricos, hallazgos cualitativos y, concluyentemente, con la evidencia estadística descritos anteriormente. En suma, estos aportes constituyen un acercamiento a un fenómeno sumamente complejo como es una teoría sobre el FDI en una población caracterizada por una discapacidad que provoca conductas de inatención e hiperactividad-impulsividad.

La principal limitación de esta investigación es que la valoración del TDAH está sujeto a la información dada por el padre o la madre del estudiante más el juicio del docente de cada asignatura y del orientador; por consiguiente, la evaluación de este trastorno es bastante arbitraria y los criterios pueden variar de una institución a otra. Ante estas circunstancias, se optó en esta investigación por el calificativo de "estudiantes reportados con conductas del TDHA", al no mediar el criterio de expertos, para identificar a los examinados del grupo focal. En el caso del grupo de referencia, se decidió emplear un atributo común a todos ellos "estudiantes sin necesidades educativas especiales", por lo que no solicitaron ningún tipo de adecuación curricular para realizar la prueba de matemática.

Referencias

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

Anastasi, A. & Urbina, S. (1998). *Tests psicológicos* (Sétima ed.). Naucalpan de Juárez, 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. *Revista de Pedagogía Bordón*, *53*, 473–483.

Andriola, W. B. (2002). *Detección del funcionamiento diferencial del ítem (DIF) en test de rendimiento. Aportaciones teóricas y metodológicas* [Versión electrónica]. Tesis doctoral, Universidad Complutense de Madrid, España. Recuperado el 22 de octubre del 2005, de http://baru.ibict.br/tede_ibict/tde_arquivos/1/TDE-2004-12-13T07:46Z-57/Público/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*, 177–188.

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

Attorresi, H., 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. Recuperado el 10 de noviembre del 2004 de: www.uv.es/psicológica/artículos2.03/7_attorresi.pdf.

Bishop, N. S., Sharairi, S., Swift, D., Wa Lei, P. & Domaleski, Ch. (2006) Alternative procedures for identifying items that are differentially difficult for ELL students. Documento presentado en Annual Meeting of the National Council on Measurement in Education. San Francisco, E.E.U.U.

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 y H. Wainer (Eds.), *Differential item functioning*, 321-335. New Jersey, EE. UU.: Lawrence Erlbaum Associates.

Carvajal, J. & Poggio, A. (2006). Studying equivalence of Spanish language versions of a large scale assessment: Differential item functioning in the cognitive and affective domain. Documento presentado en Annual Meeting of the National Council on Measurement in Education. San Francisco, Estados Unidos.

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 y H. Wainer (Eds.). *Differential item functioning,* 137–166. New Jersey, EE. UU.: Lawrence Erlbaum Associates.

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

Elosúa, P. & López, A. (1999). Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. Revista *Psicológica,* [Versión electrónica], 20, 23 - 40. Recuperado el 16 de octubre del 2004 de: www.uv.es/psicológica/artículos1.99/elosúa.

Haladyna, T. (1997). *Writing test items to evaluate higher order thinking*. MA., EE. UU.: Allyn and 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,* 2, 331-342. Recuperado el 21 de octubre del 2004, de www.um.es/anales/v15/r15-2/pdf.

Hidalgo, M. D., López, J. A. & Sánchez, J. (1997). Error tipo I y potencia de las pruebas chi-cuadrado en el estudio del funcionamiento diferencial de los ítems. En *Revista de Investigación educativa*, *15*, 149 – 168.

Longford, N. T., Holland, P. W. & Thayer, D. T. (1993). Stability of the MH D-DIF Statistics Across Populations. En P.W. Holland y H. Wainer (Eds.), *Differential item functioning,* 255–276. New Jersey, EE. UU.: Lawrence Erlbaum Associates.

Martínez, R. (2005). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid, España: Editorial Síntesis.

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

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

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

Ministerio de Educación Pública (1998). Políticas, normativa y procedimientos para el acceso a la educación de los estudiantes con necesidades educativas especiales [Reimpresión de la 1ª Ed.]. San José, Costa Rica: Autor.

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

Moreno, F. X. (2001). Análisis psicopedagógico de los alumnos de educación secundaria obligatoria con problemas de comportamiento en el contexto escolar [Versión electrónica]. Tesis Doctoral, Universitat Autònoma de Barcelona, España. Recuperado el 16 de marzo del 2006, de http://www.tdx.cbuc.es/TESIS_UAB/AVAILABLE/TDX-0726101-093527/fxmo1de1.pdf.

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

Nunnaly, J. & Bernstein, I. (1995). *Teoría psicométrica* (Tercera ed.). D.F., 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. Recuperado el 10 de noviembre del 2004 de http://www.uv.es/psicológica/artículos3.98/padilla.pdf.

Pardo, A. & Ruiz, M. A. (2002). *SPSS 11. Guía para el análisis de datos*. Madrid, España: McGraw-Hill.

Penfield, R. (2006). *DIFAS 3.01. Differential item functioning analysis system. User's Manual.* Recuperado el 2 de diciembre del 2006 de: http://education.miami.edu/facultyStaff/Faculty_Bio.asp?ID=135

Penfield, R. D. & Camilli, G. (2006). Differential Item Functioning and Item Bias. En S. Sinharay y C.R. Rao (Eds.). *Handbook of Statistics. Psychometrics*. Amsterdam, Holanda: Elsevier.

Villalobos, E. & Morales, K. (2002). *Niños con déficit de atención: Orientación a padres y docentes*. San José, Costa Rica: Editorial Universidad Estatal a Distancia.

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

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

Recibido: 7 de septiembre de 2007
Aceptado: 5 de junio de 2008