RMIE, 2016, VOL. 21, NÚM. 71, PP. 1191-1220 (ISSN: 14056666)

Investigación

ANÁLISIS DEL FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS DEL EXCALE DE MATEMÁTICAS PARA TERCERO DE SECUNDARIA

ADÁN MOISÉS GARCÍA-MEDINA / FELIPE MARTÍNEZ RIZO / GRACIELA CORDERO ARROYO

Resumen:
Este artículo presenta los resultados de un estudio de los ítems de la escala de matemáticas del Examen de la Calidad y el Logro Educativos, que se aplicó en 2012 a una muestra nacional de 27 mil 9 alumnos de tercero de secundaria en México. Se empleó un método de detección de Funcionamiento Diferencial de Ítems (DIF, por sus siglas en inglés), que parte del procedimiento original de Mantel-Haenszel con una modificación sustantiva basada en el modelo de Rasch. Se examinó la presencia de DIF a partir de dos ejes de análisis: sexo y nivel socioeconómico. De los 100 ítems de la escala de matemáticas ninguno fue identificado con DIF por sexo; en cambio, se diagnosticaron 18 con DIF severo por nivel socioeconómico.

Abstract:
This article presents the results of a study of the items from the mathematics scale of the "Examen de la Calidad y el Logro Educativos" ("Examination of Educational Attainment and Quality") that was given in 2012 to a national sample of 27,009 students in the third year of secondary school in Mexico. The method of detection involved Differential Item Functioning (DIF), starting with the original procedure of Mantel-Haenszel with a substantial modification based on the Rasch model. The presence of DIF was examined along two lines of analysis: gender and socioeconomic level. Out of the 100 items on the mathematics scale, none was identified with DIF because of gender; in contrast, 18 were diagnosed with a severe DIF due to socioeconomic level.

Palabras clave: validez, instrumentos de evaluación, evaluación del aprendizaje, análisis estadístico, México.
Keywords: validity, instruments of evaluation, evaluation of learning, statistical analysis, Mexico.

Adán Moisés García-Medina: Estudiante del doctorado en Ciencias Educativas, Universidad Autónoma de Baja California, Instituto de Investigación y Desarrollo Educativo. Km. 103 Carretera Tijuana-Ensenada, 22830, Ensenada, Baja California, México. CE: adan.moises.garcia.medina@uabc.edu.mx

Felipe Martínez Rizo: profesor-investigador de la Universidad Autónoma de Aguascalientes, Departamento de Educación. CE: felipemartinez.rizo@gmail.com

Graciela Cordero Arroyo: profesora-investigadora de la Universidad Autónoma de Baja California, Instituto de Investigación y Desarrollo Educativo. CE: gcordero@uabc.edu.mx

Introducción

En las evaluaciones masivas de logro educativo aplicadas en México a alumnos de educación básica desde inicios del siglo XXI –como los Exámenes de la Calidad y el Logro Educativos (Excale), las pruebas de la Evaluación Nacional del Logro Académico en Centros Escolares (ENLACE) o las del Programa para la Evaluación Internacional de los Estudiantes (PISA, por sus siglas en inglés)–, los resultados presentan un patrón consistente: en promedio, los alumnos de escuelas privadas consiguen los puntajes más altos, enseguida se sitúan los de urbanas públicas, primarias o secundarias generales y técnicas; después aquellos que asisten a escuelas públicas asentadas en poblaciones rurales; y, al final, los que obtienen los resultados más bajos son los estudiantes de primarias indígenas y cursos comunitarios o de telesecundarias que, en general, se ubican en las localidades más pequeñas del país y, a su vez, atienden a los alumnos con los niveles socioeconómicos más bajos (Backhoff *et al.*, 2008; Díaz y Flores, 2010; Sánchez y Andrade, 2009; Zamudio, Díaz y Lepe, 2012).

Por ejemplo, en los Excale-09, que se aplican a alumnos de tercero de secundaria, solo 12% de los estudiantes del sector privado fueron colocados en un nivel por debajo del básico[1] en la asignatura de español, sin embargo, esa situación ocurrió con 34 y 35% de los alumnos de secundarias técnicas y generales, respectivamente, e inclusive con 50% de los estudiantes de telesecundaria, que son, estos últimos, los que provienen de familias con contextos socioeconómicos más desfavorecidos. De la misma manera ocurrió con los resultados de matemáticas, donde solo 25% de los alumnos de secundarias privadas fueron clasificados en un nivel por debajo del básico, mientras que eso sucedió en 51 y 54% de los de escuelas generales y técnicas, respectivamente, y la cifra se incrementó a 62% para los de telesecundaria (Sánchez y Andrade, 2009). En el caso de primaria, aunque los porcentajes difieren, permanece el patrón que muestra cómo hay asociación entre los resultados de logro y el contexto socioeconómico (Backhoff *et al.*, 2008; Zamudio, Díaz y Lepe, 2012).

Este comportamiento en los resultados se ha interpretado como una evidencia contundente de que, reconociendo que los factores escolares influyen en el nivel de aprendizaje de los alumnos, el contexto socioeconómico de éstos y sus familias tiene un peso mayor, además de que muchas veces las condiciones de las escuelas refuerzan ese patrón regresivo, en vez de contribuir a revertirlo (Backhoff *et al.*, 2008; Blanco, 2007; Treviño *et*

al., 2013). No obstante, cabe otra explicación parcial que podría esclarecer por qué reiteradamente aparece ese patrón de resultados: que las pruebas empleadas para evaluar el logro de los alumnos de educación básica en México tengan algún tipo de sesgo derivado del origen socioeconómico y cultural de los sustentantes que ponga a algunos en desventaja.

El sesgo ocurre cuando un test o algunos de sus ítems están diseñados de tal manera que las tareas evaluativas son formuladas con un lenguaje y en un contexto que les resulta familiar a ciertos grupos poblacionales, lo que les da cierta ventaja para resolverlos o, por el contrario, que dichas tareas evaluativas a pesar de que no pretenden evaluar el desempeño sobre algún aspecto de la lengua, su formulación incluye términos desconocidos o poco comunes del contexto de determinado grupo de sustentantes, lo que ocasiona una dificultad innecesaria en los reactivos.

Un sesgo en las evaluaciones de logro educativo que se aplican en gran escala conllevaría a la generación de conclusiones erróneas sobre el desempeño educativo de ciertas subpoblaciones según su etnicidad, nivel socioeconómico, sexo, entre otras, lo que derivaría en diagnósticos imprecisos y en la eventual formulación de políticas educativas poco adecuadas a la situación que pretenden atender.

El propósito de este artículo consiste en mostrar, mediante el análisis secundario de datos, los resultados de un análisis de funcionamiento diferencial de ítems (DIF, por sus siglas en inglés) de los reactivos de matemáticas que conforman el Excale-09, que se aplicó en 2012 a alumnos de tercero de secundaria en México. Esta información, a su vez, contribuirá a la obtención de evidencias de validez con enfoque de equidad del Excale-09.

La pregunta que se pretendió responder mediante el estudio del que se deriva este artículo fue: ¿qué magnitud de DIF presentan los ítems de matemáticas del Excale que se aplicaron a estudiantes de secundaria en 2012 en México, según su sexo y nivel socioeconómico?

Referentes y justificación

La validez es el criterio más importante para evaluar la eficacia de una prueba o instrumento de medición, y se define como "el grado en el que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba y los usos que se pretende hacer de ellos" (AERA-APA-NCME, 2014:11). Esta definición, recuperada de una de las obras más influyentes en el desarrollo de instrumentos en el campo educativo y de ciencias de

la conducta –los *Standards for Educational and Psychological Tests*– aclara que la validez no es considerada como una propiedad de los instrumentos, sino que está en función de las interpretaciones y usos que se hagan de sus resultados, visión que es coincidente con Messick (1989, 1998) y Kane (2001, 2006, 2013), dos de los autores más importantes en la teoría de la validación de instrumentos.

Los *Standards* señalan que para realizar un proceso de validación adecuado, en el que se evalúen las interpretaciones realizadas a partir de las puntuaciones de una prueba en función de ciertos propósitos, se deberían obtener evidencias de cinco fuentes posibles: de contenido, de procesos de respuesta, estructura interna, de relación con otras variables y consecuencias de la evaluación (AERA-APA-NCME, 2014; AERA-NCME-APA, 1999). Además, pese a que surge a inicios de la décas de 1970 con la aparición del movimiento de los derechos civiles en Estados Unidos, una dimensión del proceso de validación que ha venido cobrando cada vez más fuerza es la relacionada con la justicia y equidad en las evaluaciones (*fairness),* entendida como la ausencia de sesgo en la medición (Gipps y Stobart, 2009; Stobart, 2005). Cualquier barrera que impida que un individuo demuestre la habilidad o constructo que se desea medir crearía un sesgo en la interpretación de los puntajes de la prueba y los usos que se hagan de ellos, lo que atentaría contra la justicia en la evaluación (Gipps y Stobart, 2010).

La equidad es un aspecto central del proceso de validación que debería ser transversal en todas las etapas de desarrollo de una prueba, desde su diseño, aplicación, interpretación de los resultados y los usos que se hagan de ellos (AERA-APA-NCME, 2014; ETS, 2014). La idea principal de la equidad en la evaluación es identificar la varianza irrelevante de constructo para modificar los ítems y eventualmente la estructura de una prueba con el fin de maximizar el desempeño de cualquier examinado, de tal manera que los puntajes sean equiparables entre los distintos sustentantes.

La validez cultural es un concepto estrechamente relacionado con el de equidad en la evaluación. Solano-Flores y Nelson-Barber definen la validez cultural como:

> […] la eficacia con la que […] la evaluación aborda la influencia socio-cultural que da forma al pensamiento del estudiante y las maneras en que los estudiantes les dan sentido a […] los reactivos y sus respuestas. Las influencias socioculturales incluyen el conjunto de valores, creencias, experiencias, patrones de

> comunicación, estilos de enseñanza y aprendizaje, epistemologías inherentes al contexto cultural de los estudiantes y las condiciones socioeconómicas que prevalecen en sus grupos culturales (citados en Basterra, Trumbull, y Solano-Flores, 2011:3).

De la definición anterior se deriva que dentro del campo de las evaluaciones de aprendizajes, la cultura influye en la forma en que los alumnos interpretan y responden a las actividades planteadas en las pruebas y que la falta de consideración de las características socioculturales de las poblaciones a las que están dirigidas puede llevar a conclusiones imprecisas e inválidas sobre su desempeño, cuestión que es de especial importancia cuando sus resultados se emplearán en decisiones de alto impacto.

La validez cultural implica que durante el diseño del test se consideren la diversidad de contextos de quienes lo responderán, como una posible fuente de varianza irrelevante al constructo que se desea medir, de tal manera que los puntajes obtenidos mediante el instrumento de evaluación no se vean sesgados por aspectos ajenos al interés de la evaluación (por ejemplo, contexto social, cultural, económico, etcétera). Para ello, es importante que durante el diseño de los test se considere su *accesibilidad*, entendida como la oportunidad sin obstáculos que "debe ofrecer una prueba a todos los sustentantes para demostrar el estado que tienen con respecto al constructo que pretende medir" (AERA-APA-NCME, 2014:49). Se parte del supuesto de que ciertos rasgos de una prueba pueden dificultar que algunos sustentantes muestren el manejo que tienen de lo que se quiere medir, por ejemplo, cuando el diseño tipográfico no permite que los sustentantes con alguna debilidad visual puedan leer los reactivos de una prueba que no está destinada a medir la capacidad visual.

En el campo de la evaluación educativa y psicológica, en la actualidad hay una clara distinción entre sesgo y funcionamiento diferencial de los ítems. El primero asociado a una dimensión ética de justicia y equidad, y el DIF como una característica meramente psicométrica. Además, este último es condición necesaria pero insuficiente para diagnosticar un sesgo en un reactivo o partes de test.

> El sesgo se refiere a la injusticia derivada de uno o varios ítems del test al comparar distintos grupos que se produce como consecuencia de la existencia de alguna característica del ítem o del contexto de aplicación del test que es irrelevante para

> el atributo medido por el ítem […] Formalmente se afirma que un determinado ítem presenta DIF cuando grupos igualmente capaces presentan una probabilidad distinta de responderlo con éxito o en una determinada dirección en función del grupo al que pertenecen (Gómez-Benito, Hidalgo y Guilera, 2010:76).

Aunque existen múltiples procedimientos para identificar el sesgo en las evaluaciones educativas, el DIF ha sido una de las herramientas más empleadas con este propósito (McNamara y Roever, 2006). El término DIF fue creado por Holland y Thayer (1988, citado en Hidalgo y Gómez-Benito, 2010) y se reconoce su presencia cuando dos individuos con el mismo nivel de habilidad tienen diferentes probabilidades de responder correctamente a un reactivo por el hecho de pertenecer a subgrupos distintos, definidos, entre otros, por estatus socioeconómico, etnicidad, contexto lingüístico, discapacidad o contexto cultural (Boone, Staver y Yale, 2014).

En México, los Excale constituyen uno de los esfuerzos más importantes para evaluar los aprendizajes de los alumnos de educación básica. Estos exámenes, diseñados y aplicados desde 2005 hasta 2014 por el Instituto Nacional para la Evaluación de la Educación (INEE) se caracterizaron por: *a)* estar alineados al currículo; *b)* ser criteriales; *c)* ser matriciales, *d)* aplicarse a una muestra representativa de alumnos; *e)* realizarse con aplicaciones rigurosamente controladas a cargo de personal ajeno a la escuela; *f)* conformarse con preguntas cerradas y abiertas; *g)* utilizar cuestionarios para alumnos, maestros, directores y padres de familia para obtener información sobre variables del contexto de la escuela misma y del hogar; *h)* realizar aplicaciones a un solo grado cada año, en un ciclo de cuatro años; e *i)* generar reportes que incluyen análisis complejos de los resultados obtenidos por los alumnos, así como de los factores del hogar y la escuela asociados con ellos (Backhoff, Peón y Sánchez, 2009; Martínez-Rizo y Santos, 2009). Para procesar las respuestas de los alumnos a las pruebas matriciales se utiliza la técnica de valores plausibles; las escalas en que se expresan los resultados y las de los constructos que se forman a partir de los cuestionarios de contexto se hacen con base en el modelo de Rasch.[2]

La preocupación por diseñar evaluaciones con enfoque de equidad es reciente. El INEE ha mostrado interés por una evaluación cuyo referente fundamental sea el derecho a la educación con la equidad como un aspecto central (INEE, 2007, 2010, 2014). Esto se refleja en el hecho de que de las principales pruebas a gran escala que se han realizado en el país, solamente para

las de Excale se han hecho análisis para identificar posibles sesgos e incluso en este caso han sido escasos y asistemáticos (Martínez-Rizo *et al.*, 2015).

Los resultados que se reportan en este trabajo permiten identificar ítems de Excale que deben ajustarse para respetar el principio de igual accesibilidad para los sustentantes y ponderar el riesgo de formular políticas que pueden impactar a estudiantes que, por pertenecer a cierta subpoblación, tienen menos probabilidades de responder correctamente reactivos que presenten sesgo severo.

A sabiendas de que para evaluar el nivel de logro de los alumnos de educación básica desde 2015 se aplican las pruebas del Plan Nacional para la Evaluación de los Aprendizajes (Planea), para este trabajo se utilizaron datos de Excale que ya estaban disponibles porque los resultados podrán ser aprovechados para el desarrollo de las nuevas pruebas, que mantienen la preocupación por la equidad y que constituye un punto central del enfoque de evaluación del INEE.

Método

El estudio es de tipo extensivo mediante el análisis secundario de datos. Consiste en un análisis DIF de los reactivos de la prueba de matemáticas para tercero de secundaria (Excale-09) empleando dos criterios de clasificación, por sexo y por nivel socioeconómico (NSE), utilizando como insumo las bases públicas que el INEE pone a disposición en su portal web. Se seleccionó la secundaria porque es el nivel de educación básica donde se aplicaron los Excale a más sustentantes.

El DIF ocurre cuando dos o más grupos de sujetos que pertenecen a cierta subpoblación, aun con la misma habilidad que se desea medir, tienen sistemáticamente diferentes probabilidades de responder correctamente a un reactivo en específico (Boone, Staver y Yale, 2014). A los grupos conformados se les identifica como grupo focal (en los que se hipotetiza la presencia de sesgo: mujeres, indígenas, etc.) y grupo de referencia (por ejemplo, hombres, no indígenas, etcétera).

Existe una considerable variedad de técnicas para detectar este fenómeno, como las derivadas de tablas de contingencias, las que utilizan la regresión ordinal logística, las que emplean el modelamiento de ecuaciones estructurales y las que se llevan a cabo a través de la teoría de respuesta al ítem (TRI) (Boone, Staver y Yale, 2014; Camilli, 2006). Sin embargo, todas las técnicas tienen un elemento en común: están orientadas a verificar que la

probabilidad de respuesta correcta o incorrecta a cada ítem en un examen sea independiente de la pertenencia de quienes lo responden a grupos de clasificación definidos por variables sociodemográficas, y homologados por sus niveles de habilidad medida (González-Montesinos y Jornet, 2012).

En esta investigación se empleó un método de detección de DIF que parte del procedimiento original de Mantel-Haenszel con una modificación sustantiva basada en el modelo de Rasch de un parámetro (Linacre y Wright, 1987). Este procedimiento modificado emplea la lógica original de Mantel-Haenszel pero incorpora las propiedades de la regresión logística en el análisis de patrones de respuesta binarios para obtener estimaciones directas del fenómeno que resultan en una mayor eficiencia y precisión en el diagnóstico de reactivos con DIF.

Se consideró que cuando el contraste en los niveles de dificultad (DIF-contrast) de un reactivo entre las subpoblaciones de interés fuese menor a 0.43 lógitos ($p<0.05$, y valor $T \geq 2$), se trataba de un reactivo con DIF insignificante; cuando fuera mayor o igual a 0.43 pero menor o igual a 0.64, el ítem era de DIF moderado; y que un reactivo presentaba un DIF severo cuando DIF-contrast >0.64 ($p<0.05$, y valor $T \geq 2$). Estos criterios son los utilizados en el modelo Rasch y son equivalentes a los que emplea el *Educational Testing Service* (ETS) en escala delta derivada de la fórmula de Mantel-Haenszel, cuyos valores son 1.5 para indicar un DIF severo y 1.0 para uno moderado (Zwick, 2012).

El DIF puede ser uniforme (en adelante DIF) o no uniforme (NUDIF). El primero se identifica cuando el DIF ocurre a lo largo de todos los niveles del continuo de habilidad de los alumnos; el segundo, cuando en grupos de estudiantes con baja habilidad, una de las subpoblaciones clasificatorias (por ejemplo, el grupo focal) tiene más probabilidades de responder correctamente a un reactivo, y en grupos con alta habilidad, otra de las subpoblaciones clasificatorias (por ejemplo, el de referencia) es la que tiene más posibilidades de contestar correctamente al mismo reactivo.

Participantes

En el ciclo escolar 2011-2012 había en México un millón 756 mil 924 alumnos de tercero de secundaria, que asistían a 36 mil 563 escuelas que funcionaban bajo cinco modalidades: generales, técnicas, telesecundarias, comunitarias y para trabajadores (Robles *et al.*, 2013). En esta investigación se trabajó con la información proveniente de una muestra de 94 mil

269 estudiantes inscritos en 3 mil 590 secundarias de México, a los que se les aplicó el Excale-09 en 2012. Como en una prueba de diseño matricial como el Excale cada alumno responde solo una parte de la prueba, para realizar los análisis DIF se utilizó la submuestra de los 27 mil 9 estudiantes que resolvieron la escala de matemáticas.

El INEE utiliza un método de muestreo bietápico para cada uno de los dominios sobre los que hace inferencias. En la primera etapa selecciona escuelas y en la segunda a los alumnos. Las primeras se eligieron de forma sistemática a partir de listados ordenados de acuerdo con la cantidad de alumnos que atienden (proporcional al tamaño); los estudiantes de cada escuela fueron seleccionados mediante un muestreo aleatorio simple (Juárez *et al.*, 2006).

Instrumentos

El INEE emplea dos instrumentos para recolectar la información que se utilizó en este trabajo:

1) *Cuestionario para el alumno.* Instrumento de 39 reactivos que recolecta información sobre el contexto social, familiar y escolar en que viven todos los estudiantes que respondieron cualquiera de las versiones o formas del Excale-09. Los ítems son de selección de respuesta.
2) *Excale-09 de matemáticas.* Test que evalúa tres subdominios: *a)* sentido numérico y pensamiento algebraico (SNPA); *b)* forma, espacio y medida (FEM); y *c)* manejo de la información (MI). La escala completa de matemáticas incluye 100 reactivos (40 de SNPA, 32 de FEM y 28 de MI). En 2012 se emplearon ocho formas o cuadernillos distintos; cada uno incluye de 37 a 39 reactivos de matemáticas. Cada ítem tiene cuatro opciones y los alumnos deben seleccionar solo una (Sánchez y Andrade, 2009).

Procedimiento

La investigación se realizó en tres etapas: cálculo del NSE, análisis DIF y NUDIF por sexo y análisis DIF y NUDIF por NSE.

Cálculo del nivel socioeconómico

El NSE es uno de los dos ejes de los análisis DIF realizados en este trabajo. A partir de la base de datos que contiene la información del cuestionario para el alumno, se seleccionaron aquellas variables (ver primeras tres

columnas del cuadro 1) que el INEE ha utilizado en otros estudios para construir índices sobre el NSE o capital cultural (cfr. Sánchez y Andrade, 2009), con la finalidad de que el procedimiento fuera comparable.

Se dividió a los estudiantes en tres estratos (bajo, medio y alto) a partir de un índice que se construyó mediante el modelamiento de crédito parcial del análisis Rasch de un parámetro utilizando el *software* Winsteps versión 3.91. En primera instancia se puso a prueba la unidimensionalidad del constructo mediante la verificación de los índices de ajuste interno (Infit) y externo (Outfit) con valores mayores a 0.8 y menores a 1.3 lógitos. Derivado de esa primera revisión, se identificó que la variable "Hacinamiento", construida a partir del cociente del reactivo AS001/AS002 no era parte de la variable latente NSE; una vez eliminada, los índices de bondad de ajuste fueron los que se presentan en las dos columnas a la derecha del cuadro 1.

Después de calibrar las variables, se procedió a calcular el índice socioeconómico a partir del nivel de habilidad del modelo Rasch. El índice se estandarizó con media de cero y desviación estándar de 1 lógito. Se crearon tres NSE, cuyos intervalos son de igual amplitud, a partir del rango de la escala que osciló de -5.23 a 4.73 lógitos.

Análisis DIF por sexo

La segunda etapa consistió en analizar la presencia de DIF y NUDIF entre hombres (grupo de referencia) y mujeres (grupo focal). Se utilizó el *software* Winsteps 3.91. En primera instancia se calcularon bajo el modelo Rasch la dificultad de cada reactivo (δ_r) y la habilidad de cada sustentante (β_s) mediante la fórmula:

$$P_i(x=1,2\ldots n|\beta_s) = e^{(\beta s - \delta r)} / (1+e^{(\beta s - \delta r)})$$

Donde $P_i(x=1|\beta_s)$ es la probabilidad de que un sustentante seleccionado aleatoriamente responda correctamente al reactivo *i*; *n* es el número de reactivos en el examen; y *e* la constante 2.71828. En una segunda instancia se calcularon los contrastes del DIF a partir de la fórmula:

$$\text{DIF-contrast} = \delta_{iH} - \delta_{iM}$$

Donde δ_{iH} es el nivel de dificultad del reactivo *i* para los hombres y δ_{iM} es el nivel de dificultad del mismo reactivo *i* para mujeres.

CUADRO 1

Variables empleadas e índices de bondad de ajuste para construir el nivel socioeconómico

Reactivo	Ítem	Categorías de respuesta	Infit MNSQ	Outfit MNSQ
Hacinamiento	AS001 Incluyéndote, ¿cuántas personas viven en tu casa? / AS002 En tu casa, ¿cuántos cuartos se utilizan para dormir?	1=1 por habitación; 2=2 por habitación; 3=3 por habitación; 4=4 o más por habitación	1.76	1.93
AS003_RE	Aproximadamente, ¿cuántos libros hay en tu casa? (no incluyas revistas, periódicos, ni tus libros escolares)	0=Ninguno; 1=1-5; 2=6-10; 3=11-15; 4=16-20; 5=21-40; 6=41-60; 7=61-80; 8=81-100; 9=101 o más	1.21	1.28
AS008_RE	¿Cuántas veces saliste con tu familia a vacacionar en los últimos dos años?	0=Ninguna; 1=1 o 2 veces; 2=3 o 4 veces; 3=5 o 6 veces; 4=7 veces o más	1.28	1.32
AS009_RE	¿Cuántos automóviles tienen en tu casa?	0=Ninguno; 1=1; 2=2; 3=3; 4=4 o más	1.17	1.27
AS011	¿Hay teléfono en tu casa?	0=No; 1=Sí	0.97	0.94
AS012	¿Hay computadora en tu casa?	0=No; 1=Sí	0.81	0.76
AS014_RE	¿Cuál es el nivel máximo de estudios que concluyó tu mamá?	0=No fue a la escuela; 1=Primaria; 2=Secundaria; 3=Bachillerato o Preparatoria; 4=Universidad	0.84	0.84
AS015_RE	¿Cuál es el nivel máximo de estudios que concluyó tu papá?	0=No fue a la escuela; 1=Primaria; 2=Secundaria; 3=Bachillerato o Preparatoria; 4=Universidad	0.83	0.82
AS016	¿Hasta qué nivel educativo te gustaría estudiar?	0=Secundaria; 1=Bachillerato o Preparatoria; 2=Carrera técnica; 3=Licenciatura o carrera universitaria; 4=Posgrado (maestría o doctorado)	1.02	1.02
AS017	¿Hasta qué nivel educativo les gustaría a tus padres que estudiaras?	0=No sé; 1=Secundaria; 2=Bachillerato o Preparatoria; 3=Carrera técnica; 4=Licenciatura o carrera universitaria; 5=Posgrado (maestría o doctorado)	0.92	0.92

Elaboración propia.

Análisis DIF por NSE

La tercera etapa consistió en analizar la presencia de DIF y NUDIF entre estudiantes de NSE bajo (grupo focal) y alto (grupo de referencia) utilizando el mismo *software* y procedimiento que en la segunda etapa.

En los dos ejes de análisis ya señalados (sexo y NSE), los análisis DIF y NUDIF se realizaron para cada una de las tres subescalas que mide el Excale-09 de matemáticas: SNPA, FEM y MI.

Posteriormente, se identificaron los ítems con sospecha de sesgo por presentar DIF o NUDIF severo, según el criterio antes mencionado: que presentasen DIF-contrast o NUDIF ≥ 0.64 lógitos. Se precisa que, en el caso de NUDIF se requiere que en uno de los terciles en que se subdividió la población según su habilidad matemática se presente DIF-contrast severo a favor del grupo focal y que en otro se registre favoreciendo al grupo de referencia o viceversa.

Resultados

A continuación se presentan los resultados de los análisis realizados para este trabajo, los cuales se organizaron de la siguiente manera: en primer lugar se muestran los análisis DIF por sexo para cada una de las tres subescalas; enseguida, los análisis NUDIF por sexo; posteriormente, los análisis DIF por NSE y, al final, los NUDIF por NSE. Como apoyo, se emplean algunas figuras que ilustran la magnitud y el sentido del funcionamiento diferencial; en el eje de las "X" se ubica cada reactivo con su clave, mientras que en el eje de las "Y" se grafica la diferencia entre el nivel de dificultad promedio de un ítem para los hombres (δ_{iH}) y para las mujeres (δ_{iM}). Así, un reactivo que no presenta DIF es aquel en el que los valores graficados para hombres y mujeres se ubican en el cero. Cuanta más distancia haya entre un valor y el cero indica un mayor funcionamiento diferencial para la subpoblación analizada. Valores positivos refieren mayor dificultad del reactivo para uno de los grupos de que se trate; por el contrario, valores negativos son indicativos de mayor facilidad para alguna de las dos subpoblaciones de análisis.

Aun cuando se señalarán los reactivos puntuales que presentaron DIF, para que el lector disponga de una visión gráfica de los ítems con un funcionamiento diferencial, a las figuras se agregan cuatro líneas de referencia que sirven para identificar el grado de DIF que presentan los reactivos. Las líneas situadas en los |0.43| lógitos sirven para distinguir entre reactivos con DIF moderado y aquellos cuyo funcionamiento diferencial es ligero

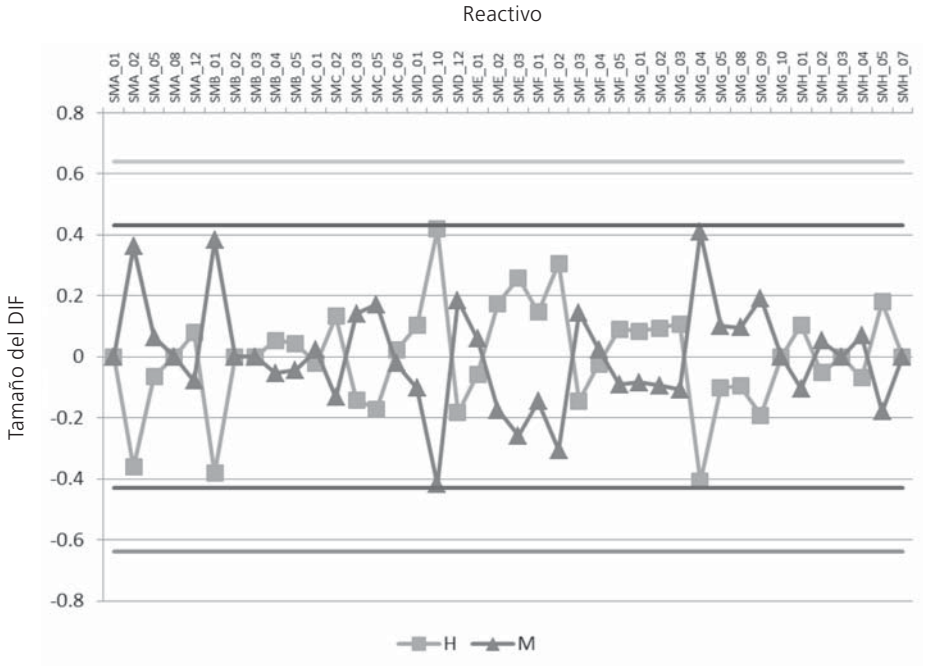
o poco significativo; por su parte, las líneas colocadas a los |0.64| lógitos permiten diferenciar entre reactivos con DIF severo y moderado. Debe advertirse que la escala del eje de las "Y" es distinta según los valores que se presenten en cada caso.

DIF por sexo

En la gráfica 1 se ilustran los resultados del análisis para la subescala de sentido numérico y pensamiento algebraico. La evidencia empírica muestra que ninguno de sus cuarenta reactivos presenta DIF entre hombres y mujeres, ni siquiera moderado. El ítem SMD_10, que fue el que tuvo mayor DIF, fue de 0.42 lógitos (p= 0.01; T= 77.89),[3] lo que todavía se considera como ligero.

GRÁFICA 1

Tamaño del DIF en la subescala SNPA del Excale-09 de matemáticas en 2012, desglosado por sexo



Elaboración propia.

En lo que corresponde a la subescala FEM, lo que muestra la evidencia es que cada uno de sus 32 reactivos está libre de DIF por sexo, lo cual implica que tanto hombres como mujeres tienen las mismas posibilidades de responder correctamente a los ítems de esta subescala, solo en función de la habilidad matemática (HABMAT) sobre el dominio FEM. Aun los reactivos con más DIF por sexo, como el SMA_07 y SMH_09, su magnitud no supera los 0.39 lógitos.

En lo referente a la subescala MI, los datos derivados de los análisis muestran que ninguno de sus 28 reactivos presenta DIF significativo por sexo, lo cual implica que tanto hombres como mujeres tienen las mismas posibilidades de responder correctamente a los ítems de esta subescala y que, más bien, lo que diferencia los puntajes que cada alumno obtiene es su HABMAT sobre el dominio MI. El reactivo SMH_06, que es el que presenta mayor DIF, es de 0.42 lógitos, valor que aún se considera aceptable.

DIF no uniforme por sexo

La presencia de NUDIF ocurre cuando al desagregar a la población de estudiantes de acuerdo con niveles de habilidad, en algunos de esos niveles los reactivos tienen más probabilidad de ser respondidos correctamente por alguno de los grupos de interés (por ejemplo, los hombres) y, en un nivel de habilidad distinto al anterior, otro grupo de interés (por ejemplo, las mujeres) es el que tiene más probabilidades de responderlo de forma correcta. Es decir, la detección de NUDIF equivale a la realización de varios análisis DIF, desagregando a los sustentantes según la habilidad evaluada.

En esta investigación, para facilitar la presentación de los resultados, se decidió dividir a los sustentantes del Excale-09 de matemáticas en terciles según su HABMAT. De esta manera, se analizó si los reactivos presentaban un funcionamiento que favoreciera a hombres o mujeres a lo largo de cada uno de los tres niveles de habilidad evaluada de los alumnos, a diferencia de los análisis previos llevados a cabo para identificar el DIF, donde no se requería desagregar a los estudiantes según su nivel de HABMAT.

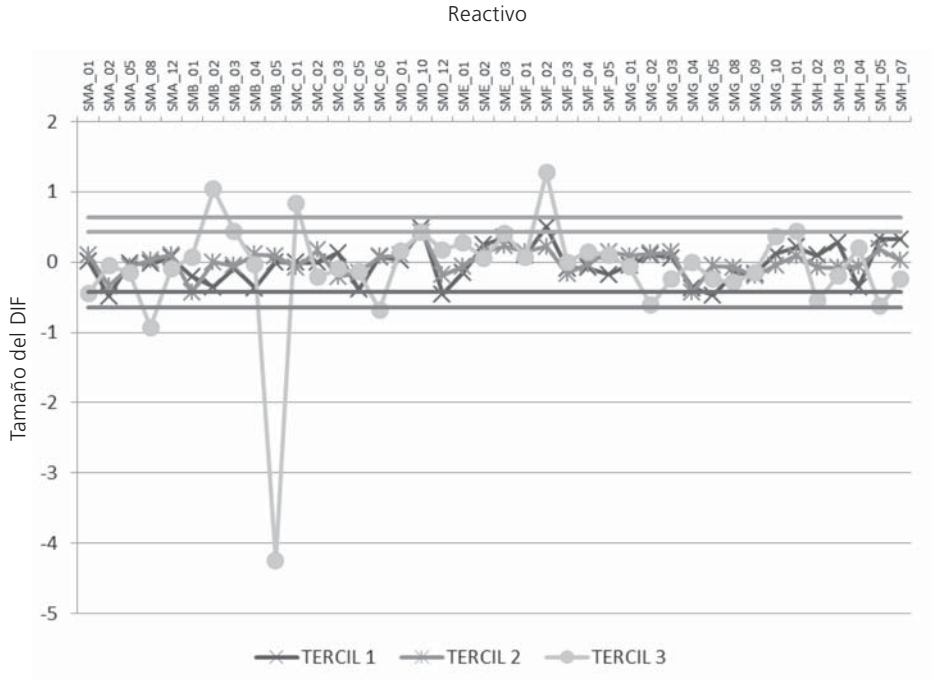
En la gráfica 2 se presentan los análisis NUDIF llevados a cabo entre hombres y mujeres con cada uno de los terciles en que se subdividió a los estudiantes según su HABMAT en la subescala SNPA. En el eje de las "X" se ubicó a cada uno de los reactivos mientras que en el de las "Y" se graficó la diferencia entre el nivel de dificultad de un reactivo para hombres ($d_{i\text{-}h}$) y mujeres ($d_{i\text{-}m}$) de cada uno de los terciles definidos en función de

su HABMAT. Valores positivos indican que el reactivo es más fácil para las mujeres; por el contrario, los negativos refieren que el ítem favorece a los hombres.

La evidencia empírica muestra que ninguno de los cuarenta reactivos de la subescala SNPA presenta NUDIF por sexo, recordemos que para que lo haya se requiere que el mismo ítem presente DIF-contrast severo a favor de un grupo en un tercil, y a favor del otro grupo en otro tercil. Sin embargo, el análisis al interior de cada tercil (que llamaremos DIF parcial) permitió encontrar algunos reactivos que presentan DIF parcial moderado (gráfica 2).

GRÁFICA 2

Tamaño del DIF no uniforme en la subescala SNPA del Excale-09 de matemáticas en 2012, desglosados por tercil de habilidad y por sexo



Elaboración propia.

En el primer tercil (33.3% de los estudiantes con menor desempeño en matemáticas), cinco reactivos tienen DIF parcial moderado, de los cuales

tres son favorables para los hombres (SMA_02, SMD_12 y SMG_05) y dos para las mujeres (SMD_10 y SMF_02).

En el segundo tercil se identificaron dos reactivos con DIF parcial moderado (SMB_01 y SMG_04), ambos con mayor probabilidad para que sean respondidos correctamente por los hombres.

En el tercil superior (33.3% de estudiantes con mayores habilidades matemáticas) se identificaron 12 reactivos con presencia de DIF parcial, de los cuales la mitad tiene DIF severo, mayor a 0.64 lógitos. Comparando a la subpoblación a la que favorecen, siete reactivos son favorables a los hombres (SMA_01, SMA_08, SMB_05, SMC_06, SMG_02, SMH_02 y SMH_05) y cinco a las mujeres (SMB_02, SMB_03, SMC_01, SMF_02 y SMH_01).

En la subescala FEM la información disponible indica que ninguno de los reactivos presenta NUDIF por sexo. De hecho, al llevar a cabo un análisis al interior de cada tercil de habilidad (DIF parcial), se halló que en los primeros dos ninguno de los 32 reactivos tiene DIF parcial severo. En el tercil superior de habilidad se identificó una docena de ítems con presencia de DIF parcial, de los cuales tres tienen DIF severo o mayor a 0.64 lógitos. Comparando a la subpoblación a la que favorecen, cinco reactivos son favorables a los hombres (SMA_04, SMB_06, SMD_03, SMD_08 y SMH_08) y los siete restantes con las mujeres (SMA_07, SMB_09, SME_05, SME_06, SMF_07, SMF_10 y SMH_10).

El análisis NUDIF de la subescala MI indica que todos sus reactivos están libres de NUDIF severo por sexo; sin embargo, se identificaron algunos con DIF parcial. Así, en el tercil inferior los reactivos SMC_08 y SMF_13 tienen un DIF parcial moderado que favorece a las mujeres. En el segundo se identificó un par de ítems (SMF_09 y SMH_06) con DIF parcial moderado que favorece a los hombres. En el tercil superior se diagnosticaron siete reactivos con DIF parcial, de los cuales cinco otorgan mayores posibilidades de responderse correctamente por los hombres (SMB_11, SMC_07, SME_10, SME_11 y SMF_12) y un par de ítems (SMB_12 y SMH_06) que son más fáciles para las mujeres.

DIF por NSE

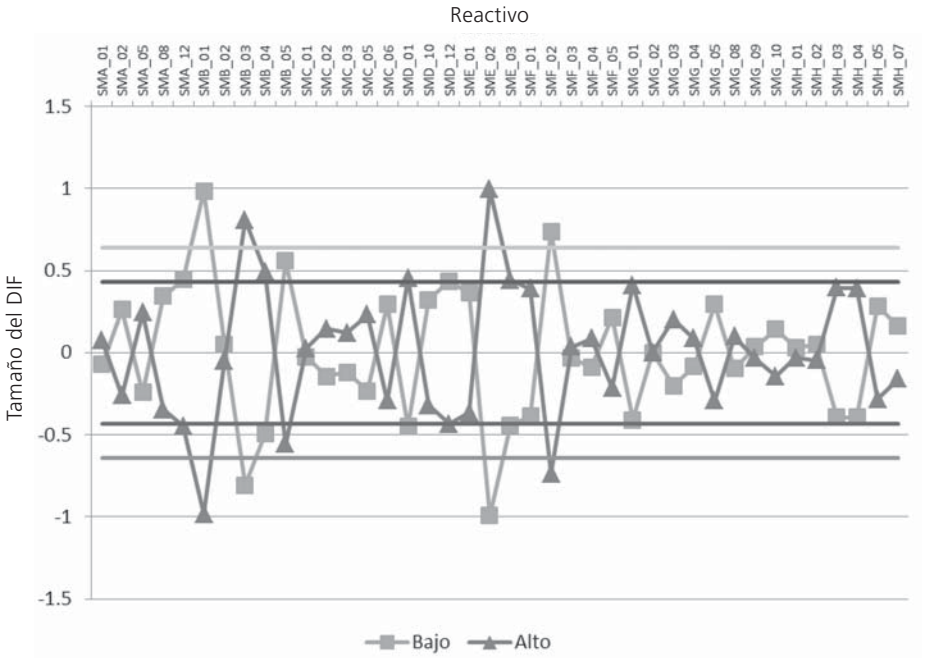
Los resultados de los análisis DIF por NSE llevados a cabo con los ítems de la subescala SNPA permitieron saber que diez de los cuarenta reactivos que la conforman presentan funcionamiento diferencial, seis de ellos de forma moderada y cuatro, severa (ver gráfica 3). De los ítems con DIF

moderado, tres tienen un funcionamiento diferencial que favorece a los alumnos de NSE alto (SMA_12, SMB_05 y SMD_12), y la otra mitad a estudiantes de NSE bajo (SMB_04, SMD_01 y SME_03); de los cuatro ítems con DIF severo, dos favorecen a estudiantes de NSE alto (SMB_01 y SMF_02) mientras que la otra mitad a alumnos que provienen de familias con NSE bajo (SMB_03 y SME_02).

En la subescala FEM se encontró que si bien la gran mayoría de sus 32 reactivos no presentan un comportamiento en los que haya una sospecha de sesgo, cinco exhiben un DIF severo entre alumnos con diferentes NSE; los ítems SMD_04, SMD_05 y SMH_11 tienen un DIF a favor de los estudiantes con NSE bajo y los reactivos SMB_09 y SMG_07 muestran un DIF que favorece a los de NSE alto. Asimismo, la evidencia muestra que tres reactivos adicionales presentaron un DIF moderado favorable hacia los alumnos con NSE alto: SMD_03, SMD_06 y SMD_09.

GRÁFICA 3

Tamaño del DIF en la subescala SNPA del Excale-09 de matemáticas en 2012, desglosado por nivel socioeconómico



Elaboración propia.

En la subescala MI se halló que la mayoría de los 28 ítems no presentan sesgo; sin embargo, se identificaron siete reactivos con DIF severo, de los cuales cuatro favorecen a los alumnos con NSE alto (SMB_11, SMC_08, SME_11 y SME_12) y los tres restantes tienen un DIF que da ventaja a los de de NSE bajo (SMC_10, SMD_11 y SME_09). Además, se encontró que el ítem SMB_12 tiene un DIF moderado a favor de los estudiantes con NSE bajo.

DIF no uniforme por NSE

Como se hizo con el análisis NUDIF por sexo, se dividió a los sustentantes del Excale-09 de matemáticas en terciles según su HABMAT. De esta manera, se analizó si los reactivos presentaban un funcionamiento que favoreciera a alumnos con NSE alto o bajo a lo largo de cada uno de los tres niveles en que se dividió la HABMAT de los alumnos.

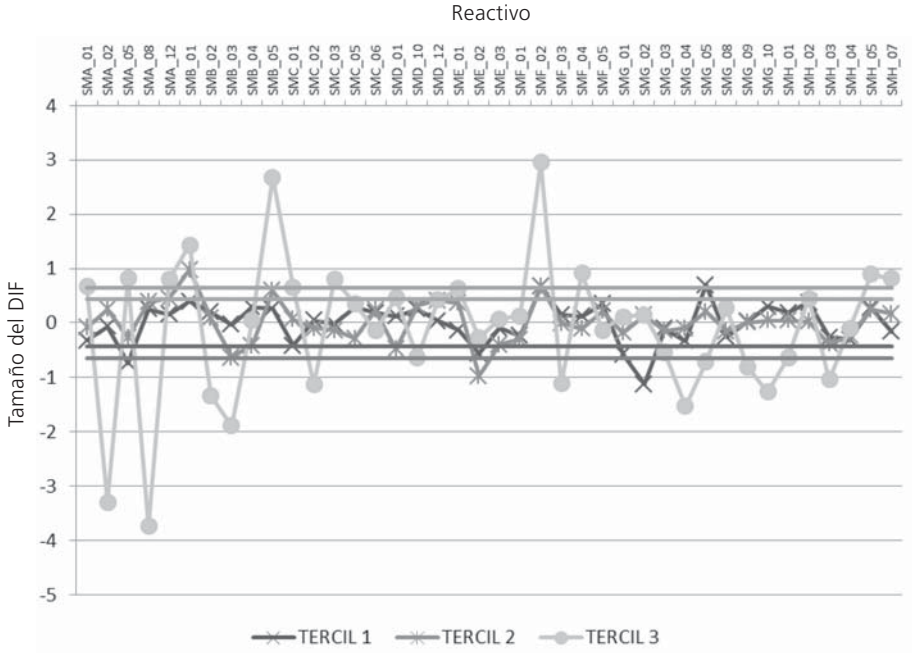
En la subescala SNPA, al llevar a cabo los análisis al interior de cada tercil de habilidad (DIF parcial), la evidencia señala que entre el subgrupo del 33% de los alumnos con menos rendimiento en matemáticas hay cuatro reactivos (SMA_05, SMG_02, SMF_02 y SMG_05) con DIF parcial severo de, al menos, 0.64 lógitos, de los cuales los dos primeros son más fáciles para los alumnos con NSE bajo y los dos últimos más fáciles para aquellos con NSE alto (ver gráfica 4). Vale recordar que aunque esta condición es necesaria no es suficiente para diagnosticar NUDIF, sino que se requiere que el DIF sea en uno de los terciles favorables hacia los alumnos de NSE alto y en otro de los terciles hacia los de NSE bajo.

En el tercil 2 de HABMAT se diagnosticaron tres reactivos con DIF parcial severo. Los ítems SMB_01 y SMF_02 son favorables para alumnos con NSE alto y el reactivo SME_02 es más fácil para estudiantes con NSE bajo.

Dentro del subgrupo conformado por la tercera parte de los alumnos de tercero de secundaria con los puntajes más altos en matemáticas se hallaron once reactivos con DIF parcial severo a favor de estudiantes con NSE alto (SMA_01, SMA_05, SMA_12, SMB_01, SMB_05, SMC_01, SMC_03, SMF_02, SMF_04, SMH_05 y SMH_07) y otros once que son más fáciles para alumnos con NSE bajo (SMA_02, SMA_08, SMB_02, SMB_03, SMC_02, SMF_03, SMG_04, SMG_05, SMG_09, SMG_10 y SMH_03). En la gráfica 4 podrá advertirse que la gran mayoría de los DIF severos ocurren en el tercil superior de HABMAT, es decir, entre los alumnos que suelen conseguir los puntajes más altos en el Excale-09 de matemáticas.

GRÁFICA 4

Tamaño del DIF no uniforme en la subescala SNPA del Excale-09 de matemáticas en 2012, desglosados por tercil de habilidad y por nivel socioeconómico



Elaboración propia.

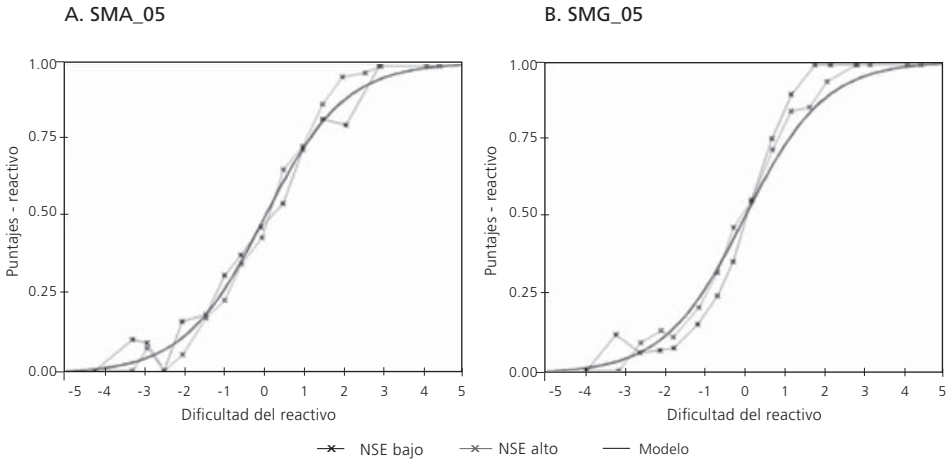
En resumen, para la subescala SNPA se identificaron dos reactivos como los únicos que cumplen el doble requisito para considerar que tiene NUDIF por NSE: SMA_05 y SMG_05. Debe subrayarse que estos ítems son los únicos en tal situación en las tres subescalas.

En el primero de ellos, dentro del subgrupo de estudiantes del primer tercil de HABMAT, los alumnos que provienen de un NSE bajo son los que tienen más probabilidades de responder correctamente al ítem; en cambio, dentro del subgrupo con los más altos puntajes en la prueba, los estudiantes con NSE alto son los que tienen más posibildades de acertar. En el reactivo SMG_05 ocurre lo opuesto: entre los estudiantes del tercil más bajo de la habilidad, los que provienen de hogares con NSE alto son los más propensos a resolver correctamente el reactivo; en cambio, al centrar la atención en el subgrupo que obtuvo los más altos puntajes, son los

estudiantes de NSE bajo quienes tienen más probabilidad de resolverlo acertadamente (gráfica 5 A y B).

GRÁFICA 5 A Y B

Curvas del funcionamiento diferencial de reactivos no uniforme, según el nivel socioeconómico de los alumnos, de reactivos SMA_05 y SMG_05



Elaboración propia.

En la subescala FEM ninguno de los 32 reactivos que la conforman presentan NUDIF en grado severo, por NSE. No obstante, al realizar DIF parciales al interior de cada uno de los terciles del continuo de la escala de HABMAT, la información disponible indica que en el primer tercil (33% de los alumnos con menor puntaje en la prueba) hay cuatro ítems con DIF parcial severo a favor de estudiantes de NSE bajo: SMD_06, SME_04, SMF_08 y SMH_11.

Dentro del segundo tercil de HABMAT se identificaron dos reactivos (SMB_09 y SMG_07) con DIF parcial severo a favor de estudiantes con NSE alto y tres que son más fáciles para de NSE bajo: SMD_04, SMD_05 y SMH_11.

En el tercil superior se encontraron nueve ítems (SMA_04, SMB_07, SMB_10, SMD_03, SMD_08, SME_05, SMG_07, SMH_08 y SMH_10) con DIF parcial severo que son más fáciles para alumnos con NSE alto y una docena de reactivos (SMA_07, SMB_08, SMD_05, SMD_06, SMD_09,

SME_06, SMF_06, SMF_07, SMF_10, SMG_06, SMH_09 y SMH_11) con DIF parcial favorable hacia los alumnos con NSE bajo. Se halló que la mayoría de ítems con DIF parcial ocurren con los estudiantes ubicados en el tercil superior de HABMAT.

En la subescala MI, compuesta por 28 ítems, no se detectó ningún reactivo con NUDIF severo. Sin embargo, al realizar análisis DIF parciales, se encontró que entre la tercera parte de estudiantes con los rendimientos más bajos en matemáticas (el tercil 1) hay dos reactivos (SMC_08 y SME_11) que tienen DIF parcial a favor de los NSE alto y uno (SMB_12) que resultó más fácil para los alumnos con NSE bajo.

En el segundo tercil se encontraron cuatro ítems (SMB_11, SMC_08, SME_11 y SME_12) que tienen un funcionamiento diferencial parcial y todos a favor de alumnos con NSE alto. Entre los estudiantes que evidenciaron los niveles más altos en la HABMAT (tercil superior) se diagnosticaron seis ítems con DIF parciales severos; cinco con DIF parcial favorable hacia estudiantes con NSE alto (SMB_11, SMC_08, SME_10, SME_11 y SMH_13) y el ítem SME_12 que resultó más fácil para los alumnos provenientes de familias con NSE bajo.

Conclusiones

Aplicando los criterios señalados en el apartado metodológico, los análisis DIF y NUDIF realizados con los cien reactivos del Excale-09 de matemáticas permiten concluir que ninguno de ellos presentó DIF o NUDIF por sexo, en tanto que se identificaron 18 ítems con sospecha de sesgo entre alumnos de distinto NSE.

El cuadro 2 presenta los 18 ítems con sospecha de sesgo. De la subescala SNPA se identificaron cuatro reactivos con DIF severo (SMB_01, SMB_03, SME_02, SMF_02) y dos con NUDIF severo (SMA_05, SMG_05). En la subescala FEM se hallaron cinco ítems con DIF severo (SMB_09, SMD_04, SMD_05, SMG_07 y SMH_11) y ninguno con NUDIF. De la subescala MI se encontraron siete reactivos con DIF severo (SMB_11, SMC_08, SMC_10, SMD_11, SME_09, SME_11 y SME_12) y ninguno con NUDIF.

La primera columna del cuadro 2 presenta las tres subescalas de matemáticas y la segunda los ítems sospechosos de sesgo. En la columna DIF y las tres de NUDIF se señala, con los símbolos que se explican al pie del cuadro, el nivel de sesgo moderado o severo, y su sentido favorable a una u otra de las subpoblaciones consideradas. Puede observarse que el DIF no

es predominante para cierto subgrupo poblacional, sino que aproximadamente la mitad de los reactivos favorecen a los alumnos de NSE alto y la otra mitad a los de NSE bajo.

CUADRO 2

Resumen de reactivos con DIF y NUDIF por nivel socioeconómico del Excale-09 de matemáticas en 2012

Subescala	Ítem	DIF	NUDIF		
			Tercil 1	*Tercil 2*	*Tercil 3*
SNPA	SMA_05		--		++
	SMB_01	++		++	++
	SMB_03	--		-	--
	SME_02	--	-	--	
	SMF_02	++	++	++	++
	SMG_05		++		--
FEM	SMB_09	++	+	++	
	SMD_04	--		--	
	SMD_05	--		--	--
	SMG_07	++		++	++
	SMH_11	--	--	--	--
MI	SMB_11	++	+	++	++
	SMC_08	++	++	++	++
	SMC_10	--		-	
	SMD_11	--		-	
	SME_09	--	-	-	
	SME_11	++	++	++	++
	SME_12	++		++	--

Notas: + DIF moderado a favor de NSE alto; ++ DIF severo a favor de NSE alto
- DIF moderado a favor de NSE bajo; -- DIF severo a favor de NSE bajo

Elaboración propia.

Si bien el análisis DIF es una técnica que se ha utilizado desde hace más de tres décadas con un considerable desarrollo en la psicología y ciencias de la salud (Gómez-Benito, Hidalgo y Guilera, 2010), en el campo educativo son pocas las evaluaciones de logro aplicadas en gran escala que emplean DIF desde el diseño, y aún menos las que reportan sus resultados, lo que dificulta disponer de un punto de comparación con la información aquí presentada.

La *National Assessment of Educational Progress* (NAEP), una de las evaluaciones nacionales más antiguas (se ha aplicado desde 1969) para medir el rendimiento de los alumnos en Estados Unidos en diferentes dominios como matemáticas, lectura, ciencias, escritura, historia, entre otras, es de las escasas pruebas de corte nacional que publican sus resultados de DIF. La NAEP de matemáticas se aplica bajo dos esquemas: transversal y longitudinal. En el primero se pretende generar diagnósticos sobre determinado ciclo escolar y se aplica en los grados 4, 8 y 12 de la educación básica. Bajo el esquema longitudinal se pretende valorar el progreso educativo de ciertas cohortes de alumnos y se aplica a los de 9, 13 y 17 años. Los análisis DIF de la NAEP se llevan a cabo para identificar posibles fuentes de sesgo bajo tres ejes de análisis: *a)* sexo (hombres-mujeres), *b)* raza (afroamericanos-blancos) y *c)* raza (hispanos-blancos).

En las NAEP de matemáticas, si bien hay algunas versiones en las que ninguno de sus ítems presenta DIF severo entre hombres y mujeres, afroamericanos, hispanos y blancos, hay otras versiones, principalmente las utilizadas en el esquema longitudinal, donde el 7 u 8% de sus reactivos fueron diagnosticados con un fuerte DIF en al menos una de las subpoblaciones analizadas (cuadro 3).

En América Latina, el sistema educativo chileno, uno de los más avanzados en la región junto con México en materia de evaluación educativa, también analiza el DIF de los reactivos del Sistema de Medición de Logros de Aprendizaje (SIMCE) a partir de dos ejes: sexo y ruralidad; sin embargo, en su documentación técnica o en la literatura especializada no se presenta la cantidad de sus ítems con funcionamiento diferencial.

En México el INEE, como instancia responsable del diseño y la aplicación de las pruebas Excale, ha manifestado interés por atender lo relativo a equidad en las evaluaciones que realiza (INEE, 2015; Ruiz *et al.*, 2015). El desarrollo de las pruebas incluía procedimientos para ello en el diseño de los ítems, pero no con el detalle necesario para atender suficientemente un

aspecto particularmente complejo, según señala la literatura especializada (Martínez-Rizo *et al.*, 2015).

CUADRO 3

Porcentaje de reactivos con DIF severo en las pruebas de matemáticas de la NAEP, desglosados tipo de prueba, año de aplicación y subpoblación que presenta DIF

Tipo de prueba	Año de aplicación	Grado Edad	Total de reactivos	Reactivos con DIF severo			Total	%
				Mujeres-hombres	*Afroamer.-blancos*	*Hispanos-blancos*		
Transversal	2007	4°	164	0	2	4	6	4
		8°	167	0	0	0	0	0
	2005	4°	168	0	1	2	3	2
		8°	177	0	0	0	0	0
		12°	180	2	0	1	3	2
	2003	4°	179	0	1	2	3	2
		8°	196	1	1	2	4	2
	2000	4°	26	0	0	0	0	0
		8°	27	0	0	0	0	0
		12°	27	0	0	0	0	0
Longitudinal	2008	9 años	135	0	5	5	10	7
		13 años	153	2	3	1	6	4
		17 años	155	12	1	0	13	8
	2004	9 años	140	1	0	1	2	1
		13 años	166	7	3	4	14	8
		17 años	162	5	0	8	13	8

Elaboración propia a partir de documentación técnica de la NAEP. Disponible en: https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_avoidviolat_results.aspx

Las orientaciones que se dan a los comités de validación que analizan los reactivos de Excale indican que sus integrantes deben revisar el contenido curricular a que se refiere el ítem, los posibles problemas técnicos de su construcción, el sesgo de género y cultural. Este último aspecto es sobre el que se solicitan menos indicadores de invalidez, que se centran sobre todo

en cuestiones de género, el uso de vocabulario potencialmente agresivo o denigrante y el uso de regionalismos (Backhoff *et al.*, 2006).

La atención especial que se ha dado a posibles sesgos de género es consistente con los resultados de esta investigación, en el sentido de que ninguno de los ítems del Excale-09 de matemáticas fue diagnosticado con DIF o NUDIF severo entre hombres y mujeres, lo que resulta muy adecuado si se considera que aun cuando en los sistemas de evaluación de logro con más desarrollo, como las NAEP, todavía algunos de sus ítems presentan DIF por sexo, reactivos que luego son corregidos para solventar este problema. Por el contrario, las NAEP tienen una menor proporción de reactivos con DIF severo por raza que los Excale en lo referente a NSE, y aunque estas dos variables no son estrictamente comparables, es plausible este balance porque en Estados Unidos existe una fuerte asociación entre ellas (ver cuadro 3).

En cuanto al posible sesgo por nivel socioeconómico, para confirmar que los 18 reactivos señalados en el cuadro 2 tienen realmente sesgo deberá analizarse cada uno con otros estudios, por ejemplo mediante protocolos verbales con sujetos de las subpoblaciones consideradas o con jueces de expertos que confirmen la presencia de elementos dentro de un reactivo que dificulten que alumnos de cierto NSE utilicen todo su talento para responderlo.

Una de las hipótesis que explicaría la presencia de DIF en los reactivos de matemáticas es la inclusión de términos (no relacionados con el desempeño en matemáticas) que son desconocidos para los subgrupos de estudiantes poblaciones donde se presentó el funcionamiento diferencial o bien, que el contexto donde se sitúa el problema a resolver es ajeno para los sustentantes. Estos subgrupos pueden ser de NSE bajo o alto.

Los resultados parciales de un estudio que está en marcha a través de entrevistas cognitivas para analizar en profundidad los 18 reactivos con DIF, ya han permitido identificar cómo la presencia de ciertos términos que buscan evaluar la competencia de los alumnos para resolver problemas aditivos (por ejemplo, "descendió" en lugar de emplear una palabra más común como "bajó"), ocasionan que estudiantes de NSE bajo, al desconocer dicha palabra o comprenderla con un significado opuesto, responden de manera errónea, a pesar de contar con la competencia matemática suficiente.

Para atender de la mejor manera posible la exigencia de equidad, la literatura especializada actual recomienda que para desarrollar pruebas

de rendimiento en gran escala, la atención a la diversidad se considere desde su diseño (Gipps y Stobard, 2010; Solano-Flores, 2011), como un elemento a atender durante las diferentes fases de la elaboración de pruebas y que implica, entre otras cuestiones, incluir en los diferentes comités encargados de diseñar y validar los ítems a especialistas que ayuden a identificar potenciales fuentes de sesgo, tales como antropólogos, lingüistas, sociólogos y expertos en determinados contextos socioculturales. Todo ello en el marco de una metodología de elaboración de instrumentos que va generando cada vez más adeptos, conocida como *diseño universal,* que se aplica a "aquellas evaluaciones diseñadas y desarrolladas desde el inicio para favorecer la participación del mayor número de estudiantes, y que los resultados lleven a inferencias válidas acerca del desempeño de todos los estudiantes" (Thompson, Johnstone y Thurlow, 2002:5), con lo que se promueva "el acceso a la mayor cantidad de sustentantes, sin importar su raza, origen étnico, género, estado socioeconómico, discapacidad, lengua o contexto cultural" (AERA-APA-NCME, 2014:50).

Los análisis realizados en esta investigación son una herramienta psicométrica que puede advertir a los diseñadores de evaluaciones estandarizadas, y a quienes usan sus resultados, sobre la posible presencia de sesgo. Si se considera que dentro de la región, incluso las pruebas del SIMCE del sistema chileno no consideran DIF por NSE, realizar estos análisis con los Excale es un paso necesario para generar pruebas estandarizadas más justas pero, al mismo tiempo, es un avance insuficiente que hace ver cómo este tipo de pruebas requieren de mecanismos que aseguren que la atención a la diversidad es un elemento que se considera desde sus fases iniciales de su diseño hasta la manera en que se utilizan sus resultados.

Una segunda parte de este trabajo, que está en curso, incluirá análisis de los 18 reactivos identificados con protocolos verbales con jóvenes de distinto NSE y expertos, con lo cual se aportarán evidencias de validez del Excale-09, lo que se espera contribuya a la realización de evaluaciones más justas para los alumnos mexicanos, sin importar su nivel socioeconómico. No cuidar sistemáticamente la validez cultural en las evaluaciones en gran escala (no solo las de alumnos, sino también las que se aplican a docentes) daría lugar a decisiones de política basadas en resultados sesgados, que podrían favorecer o perjudicar de manera injusta a ciertos grupos sociales clasificados por nivel socioeconómico, cultural, lingüístico y tipo de localidad donde residen.

Notas

[1] Los Excale sitúan a los alumnos en cuatro niveles de logro: avanzado, medio, básico y por debajo del básico. "En términos generales, el nivel por debajo del básico indica carencias importantes en el dominio de los contenidos curriculares, lo cual señala dificultades serias para continuar aprendiendo. El nivel básico representa un dominio elemental para progresar. El nivel medio significa un dominio sustancial de los contenidos, lo cual manifiesta un buen aprovechamiento de los contenidos prescritos en el currículo. Por último, el nivel avanzado implica un dominio óptimo de los contenidos, lo cual evidencia el máximo aprovechamiento respecto de lo previsto" (Sánchez y Andrade, 2009:14).

[2] Para aquellos interesados en conocer con detalle el proceso de diseño y calificación de los Excale, se puede consultar la documentación técnica en: http://www.inee.edu.mx/index.php/proyectos/excale/excale-documentos-tecnicos

[3] Para aligerar la lectura, en adelante se omitirán los valores de p y T cuando $p < 0.05$ y $T \geq 2$.

Referencias

AERA-APA-NCME (2014). *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association- American Psychological Association-National Council on Measurement in Education.

AERA-NCME-APA (1999). *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association-National Council on Measurement in Education-American Psychological Association.

Backhoff, Eduardo; Peón, Margarita; Sánchez, Andrés y Andrade, Edgar (2006). *Excale. Manual técnico para la validación de reactivos*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Backhoff, Eduardo; Bouzas, Arturo; González-Montesinos, Manuel; Andrade, Edgar y Hernández, Eduardo (2008). *Factores asociados al aprendizaje de estudiantes de 3º de primaria en México*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Backhoff, Eduardo; Peón, Margarita y Sánchez, Andrés (2009). *Manual técnico para el diseño de Exámenes de la Calidad y el Logro Educativos*, 2ª ed., Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Basterra, María del Rosario; Trumbull, Elise y Solano-Flores, Guillermo (2011). *Cultural Validity in Assessment*, Nueva York/Londres: Routledge.

Blanco, Emilio (2007). *Eficacia escolar en México: factores escolares asociados a los aprendizajes en la educación primaria*, Ciudad de México: FLACSO-Sede México. Disponible en: http://flacsoandes.edu.ec/dspace/handle/10469/1247 (consultado el 5 de noviembre de 2014).

Boone, William; Staver, John y Yale, Melissa (2014). "Differential Item Functioning", en *Rasch analysis in the human sciences*, Dordrecht: Springer Netherlands, pp. 273-297. Disponible en: http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-94-007-6857-4_13 (consultado el 29 de enero de 2015).

Camilli, Gregory (2006). "Test Fairness", en R. L. Brennan (ed.), *Educational measurement*, 4ª ed., Washington, D.C: ACE-National Council on Measurement in Education, pp. 221–256.

Díaz, María Antonieta y Flores, Gustavo (2010). *México en PISA 2009*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

ETS (2014). *ETS Standards for quality and fairness,* Princeton, NJ: Educational Testing Service.

Gipps, Caroline y Stobart, Gordon (2009). "Fairness in assessment", en C. Wyatt-Smith y J. Cumming (eds.), *Educational assessment in the 21st Century*, Dordrecht: Springer Netherlands, pp. 105–118. Disponible en: http://link.springer.com/10.1007/978-1-4020-9964-9.

Gipps, Caroline y Stobart, Gordon (2010). "Fairness", en P. Peterson, E. Baker y B. MacGaw (eds.), *International Encyclopedia of Education*, 3ª ed., Oxford: Elsevier-Academic Press, pp. 56-60.

Gómez-Benito, Juana; Hidalgo, M. Dolores y Guilera, Georgina (2010). "El sesgo de los instrumentos de medición. Tests justos", *Papeles del Psicólogo,* vol. 31, núm. 1, pp. 75–84. Disponible en: http://www.redalyc.org/pdf/778/77812441008.pdf.

González-Montesinos, Manuel y Jornet, Jesús (2012). *Procedimientos para la detección del Funcionamiento Diferencial de Reactivos (DIF)*, documento preparado para el INEE, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Hidalgo, M. Dolores y Gómez-Benito, Juana (2010). "Differential Item Functioning", en P. Peterson, E. Baker y B. MacGaw (eds.), *International Encyclopedia of Education,* 3ª ed., Oxford: Elsevier-Academic Press. Disponible en: http://www.sciencedirect.com/science/article/pii/B9780080448947002426.

INEE (2007). *La educación para poblaciones en contextos vulnerables. Informe anual 2007*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

INEE (2010). *El derecho a la educación en México. Informe 2009*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

INEE (2014). *El derecho a una educación de calidad. Informe 2014*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

INEE (2015). *Primeros resultados de la Evaluación de Condiciones Básicas para la Enseñanza y el Aprendizaje (ECEA) 2014. Primaria*, Textos de divulgación, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Juárez, Elizabeth; Ramírez, Ricardo y Rodríguez, Gustavo (2006). *Manual técnico para el muestreo poblacional. Excale*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Kane, Michael (2001). "Current concerns in validity theory", *Journal of Educational Measurement* vol. 38, núm. 4, pp. 319–342. Disponible en: http://www.jstor.org/stable/1435453.

Kane, Michael (2006). "Validation", en R. L. Brennan (ed.), *Educational measurement,* 4ª ed, Washington, D.C: ACE-NCME, pp. 17–64.

Kane, Michael (2013). "Validating the interpretations and uses of test scores", *Journal of Educational Measurement,* vol. 50, núm. 1, pp. 1-73. Disponible en: http://onlinelibrary.wiley.com/doi/10.1111/jedm.12000/full (consultado el 9 de octubre de 2013).

Linacre, John y Wright, Benjamin (1987). *Item Bias: Mantel-Haenszel and the Rasch Model. Memorandum No. 39*, reporte de investigación, Finnish Association of Mathematics

and Science Education Research. Disponible en: http://eric.ed.gov/?id=ED281859 (consultado el 30 de noviembre de 2014).

Martínez-Rizo, Felipe (coord.) (2015). *Las pruebas ENLACE y Excale. Un estudio de validación*. Cuadernos de Investigación, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Martínez-Rizo, Felipe y Santos, Annette (2009) "Evaluación e investigación educativa", en A. Alba (ed.), *¿Qué dice la investigación educativa?*, Ciudad de México: Consejo Mexicano de Investigación Educativa, pp. 265-304. Disponible en: http://www.riieeme.mx/docs/docs%202013/Martinez%20Rizo_2009%20Evaluacion%20educativa.pdf (consultado el 3 de diciembre de 2014).

McNamara, Tim y Roever, Carsten (2006). "Psychometric approaches to fairness: Bias and DIF", *Language Learning,* vol. 56, núm. S2, pp. 81-128. Disponible en: http://dx.doi.org/10.1111/j.1467-9922.2006.00381.x.

Messick, Samuel (1989). "Validity", en R. L. Brennan (ed.), *Educational measurement,* 3ª ed., Nueva York: ACE-National Council on Measurement in Education, pp. 13-103.

Messick, Samuel (1998). "Test validity: A matter of consequence", *Social Indicators Research,* vol. 45, núms. 1-3, pp. 35-44. Disponible en: http://dx.doi.org/10.1023/A%3A1006964925094.

Robles, Héctor; Hernández, Juan Manuel; Zendejas, Laura y Pérez, Mónica (2013). *Panorama educativo de México 2012. Indicadores del sistema educativo nacional. Educación básica y media superior*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Ruiz Cuéllar, Guadalupe; Pérez, María Guadalupe; Langford, Patricia y García-Medina, Adán Moisés (2015). *Atención a la diversidad en evaluaciones educativas externas. Muestra de prácticas internacionales*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Sánchez, Andrés y Andrade, Edgar (2009). *El aprendizaje en tercero de secundaria en México. Informe sobre los resultados del Excale 09, aplicación 2008 Español, Matemáticas, Biología y Formación cívica y ética*, Ciudad de México: Instituto Nacional para la Evaluación de la Educación.

Solano-Flores, G. (2011). "Assessing the cultural validity of assessment practices", en *Cultural validity in assessment,* Nueva York: Routledge, pp. 3-21.

Stobart, Gordon (2005). "Fairness in multicultural assessment systems", *Assessment in Education: Principles, Policy & Practice,* vol. 12, núm. 3, pp. 275-287. Disponible en: http://dx.doi.org/10.1080/09695940500337249 (consultado el 12 de diciembre de 2014).

Thompson, Sandra; Johnstone, Christopher y Thurlow, Martha (2002). *Universal design applied to large scale assessments*, National Center on Educational Outcomes. Disponible en: https://osepideasthatwork.org/Toolkit/pdf/Universal_Design_LSA.pdf (accedido el 27 de noviembre de 2015).

Treviño, Ernesto; Place, Katherine; Gempp, Rene y Donoso, Francisco (2013). *Análisis de los factores latentes y su vínculo con los resultados académicos de los niños*, Santiago, Chile:

OREALC-Unesco. Disponible en: http://www.unesco.org/new/fileadmin/MULTIMEDIA/FIELD/Santiago/pdf/factores-asociados-al-aprendizaje-en-el-serce.pdf (consultado el 10 de diciembre 2014).

Zamudio, Celia; Díaz, Celia y Lepe, Enrique (2012). *El aprendizaje de los contenidos curriculares de español un análisis de los resultados de los Exámenes de la Calidad y el Logro Educativos (Excale 03, 06 y 09)*, Ciuda de México: Instituto Nacional para la Evaluación de la Educación.

Zwick, Rebecca (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement.* ETS Research Report núm. 8. Disponible en: http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2012.tb02290.x/abstract (accedido el 9 de julio de 2015).

Artículo recibido: 19 de enero de 2016
Dictaminado: 14 de mayo de 2016
Segunda versión: 9 de junio de 2016
Aceptado: 20 de junio 2016