

EL SESGO DE LOS INSTRUMENTOS DE MEDICIÓN. TESTS JUSTOS

Juana Gómez-Benito¹, M. Dolores Hidalgo² y Georgina Guilera¹

¹Universidad de Barcelona. ²Universidad de Murcia

Las evaluaciones psicológicas deben garantizar la equidad y validez de las interpretaciones y decisiones adoptadas a partir de las mismas. Para ello es necesario la utilización de instrumentos libres de sesgo, y capaces de evaluar necesidades personales y sociales de individuos con diferentes características. El estudio sobre el posible sesgo de los tests, o de parte de sus ítems, ha ocupado un lugar relevante en la investigación psicométrica de los últimos 30 años y es previsible que siga constituyendo un importante foco de interés para los profesionales e investigadores implicados en la evaluación mediante el uso de los tests. Este trabajo pretende abordar esta perspectiva ofreciendo al psicólogo aplicado unas directrices y un bagaje de conocimientos sobre los conceptos de sesgo, funcionamiento diferencial e impacto, los procedimientos de detección de ítems o tests sesgados y la evaluación de sus posibles causas para, en conjunto, mejorar la validez de las mediciones psicológicas.

Palabras clave: Sesgo, Funcionamiento diferencial del ítem, Procedimientos de detección, Tests justos, Validez.

Psychological assessment must ensure the equity and validity of interpretations and of any decisions taken as a result of them. That is it necessary the use of bias-free assessment instruments those are capable of evaluating the personal and social needs of individuals with different characteristics. The study about the possible bias of tests, or some of their items, has had great relevance in psychometric research for the last 30 years and it will probably continue to be an important focus of interest for professionals and researchers involved in psychological and educational testing. The aim of this paper is providing to the applied psychologist the background about bias, differential functioning and impact concepts, item or tests bias detection procedures and evaluation of its possible causes and, therefore, for improving the validity of psychological measurement.

Key words: Bias, Differential Item Functioning (DIF), Detection procedures, Fair tests, Validity.

Los tests constituyen uno de los instrumentos de medida estandarizados más empleados en las ciencias sociales y de la salud, especialmente en psicología y educación. No hay que olvidar que un test se administra con un objetivo concreto, generalmente para tomar decisiones que en la mayoría de ocasiones son relevantes para la vida del individuo receptor. Así por ejemplo, en España se emplean tests para entrar a los cuerpos de seguridad, conseguir un puesto de trabajo, superar una materia en la universidad, formar parte de un programa de intervención, entre otros. Por lo tanto, es de extrema importancia que los profesionales que emplean este tipo de instrumentos se cercioren de que garantizan la igualdad de oportunidades y el tratamiento equitativo de los individuos a los que se les administra el test en cuestión, en otras palabras, que el test sea justo en las decisiones que de él se derivan.

Pero, ¿cuándo podemos afirmar que un test es justo?

Decidir hasta qué punto un test está siendo justo en su

medición no es tarea fácil. Aspectos como el contexto sociocultural, el proceso de construcción y/o adaptación, las condiciones de aplicación, la interpretación de las puntuaciones y el grado de formación del profesional (Muñiz y Hambleton, 1996) pueden ocasionar que el test sea injusto en su aplicación. Como afirman estos autores, la mayoría de los problemas en torno a los tests provienen de su uso inadecuado, más que del test en sí, de su construcción, o de sus propiedades técnicas. Por lo tanto, asumiendo que las dos primeras cuestiones están solventadas, el interés se traslada a las propiedades técnicas o psicométricas del test.

SESGO, IMPACTO Y DIF

En este contexto, la presencia de un posible sesgo en los ítems que componen el test es una preocupación central en la evaluación de la validez de los instrumentos de medida, entendiendo por validez el grado en que la evidencia empírica y el razonamiento teórico apoyan la adecuación e idoneidad de las interpretaciones basadas en las puntuaciones de acuerdo con los usos propuestos por el test (Messick, 1989; Prieto y Delgado, 2010). Así pues, cuando afirmamos que un test determinado es váli-

Correspondencia: Juana Gómez-Benito. Dpto. Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Barcelona. Paseo Valle Hebrón, 171. 08035-Barcelona. España. E-mail: juanagomez@ub.edu

do, lo que realmente estamos diciendo es que la puntuación obtenida tiene un significado específico, asumiendo que este significado es el mismo en los distintos grupos para los cuales el test ha sido validado. No obstante, para garantizar que una puntuación de un test presenta el mismo significado en diversos grupos, se requieren numerosos estudios que evalúen distintas evidencias de la validez del test (APA, AERA, y NCME, 1999). La existencia de sesgo en los instrumentos de medida psicológicos puede representar una seria amenaza contra la validez de dichos instrumentos en los que algunos de sus ítems están beneficiando a ciertos grupos de la población en detrimento de otros de igual nivel en el rasgo que interesa medir. De modo complementario, el hecho de que no haya sesgo de los ítems representa una evidencia del grado de generalización de las interpretaciones basadas en las puntuaciones del tests para distintos subgrupos de una o varias poblaciones.

El tema del sesgo ha acaparado la atención de los investigadores y profesionales, especialmente desde la polémica generada por los estudios de Jensen (1969, 1980). Este autor consideró que la inteligencia era hereditaria y que, por tanto, las diferencias que se observaban entre grupos raciales eran atribuibles a la genética. Evidentemente esta afirmación activó efusivas discusiones entre genetistas y ambientalistas. Estos últimos defendían que la explicación de las diferencias entre los grupos había que buscarla en el posible sesgo cultural de los tests de inteligencia. En ese momento, el papel de los psicómetras se centró en averiguar hasta qué punto las diferencias entre grupos eran debidas a características reales de los individuos de cada grupo o a artefactos generados por el propio instrumento. Este debate generó un nuevo conflicto semántico: ¿sesgo cultural o propiedades psicométricas distintas?

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, mientras que el segundo hace referencia únicamente a las características psicométricas del ítem. En la actualidad se ha llegado al acuerdo que el término sesgo asume que se conocen o se investigan las causas por las cuales determinados ítems presentan un comportamiento diferencial en función de ciertas variables, cuando en la mayoría de estudios lo único que se puede inferir es si existen diferencias en los resultados conseguidos por distintos individuos igualmente capaces. El término adecua-

do para este último tipo de resultados, que únicamente hace referencia a las propiedades psicométricas, es funcionamiento diferencial del ítem (Differential Item Functioning, DIF), denominación adoptada a raíz de la publicación de Holland y Thayer (1988) con la finalidad de distinguir entre ambos conceptos.

Formalmente se afirma que un determinado ítem presenta DIF si a nivel psicométrico se comporta diferencialmente para diversos grupos, es decir, el DIF indica una diferencia del funcionamiento del ítem (o test) entre grupos comparables de examinados, entendiendo por comparables aquellos grupos que han sido igualados respecto al constructo o rasgo medido por el test (Potenza y Dorans, 1995). En otras palabras, un 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. En la terminología propia del DIF se denomina *grupo focal* al conjunto de individuos, generalmente minoritario, que representa el foco de interés del estudio y que normalmente es el grupo desaventajado, mientras que el *grupo de referencia*, generalmente mayoritario, se refiere a un grupo de individuos estándar respecto al cual se compara el grupo focal. Sin embargo, el hecho que un instrumento de medida obtenga resultados sistemáticamente inferiores en un grupo en comparación a otro no necesariamente implica la presencia de DIF, sino que pueden existir diferencias reales entre los grupos en el rasgo medido por el test en cuestión. En este caso se habla de impacto (Camilli y Shepard, 1994) o diferencias válidas (van de Vijver y Leung, 1997).

Una vez aclarada la diferencia entre sesgo, DIF e impacto, imaginemos que estamos estudiando un ítem potencialmente sesgado en contra de un grupo minoritario. ¿Cómo podemos evaluar la presencia de DIF? La lógica probablemente nos llevaría a comparar directamente las puntuaciones del ítem del grupo minoritario frente al resto de examinados, y si se observasen diferencias diríamos que el ítem está siendo injusto con uno de los grupos. Sin embargo, no podemos tener la certeza de si las diferencias provienen del sesgo del ítem o realmente el nivel de habilidad de un grupo y otro son distintos. El concepto de DIF pretende abordar esta cuestión, por este motivo los análisis de DIF comparan las respuestas al ítem entre los grupos únicamente cuando éstos han sido igualados en el nivel de habilidad o del rasgo medido mediante un criterio de igualación. En este sentido, es imprescindible disponer de un criterio libre de sesgo; no obstante, en la mayoría de situaciones la única evidencia

empírica de equiparación o igualación de que se dispone es el propio test (generalmente la puntuación total), que se encuentra contaminada por la presencia de ítems con DIF y que forman parte del criterio juntamente con los ítems sin DIF. Por lo tanto, un problema endémico a los métodos de detección del DIF reside en que adolece de una cierta circularidad en su forma de proceder ya que el ítem estudiado también contribuye a la definición de la variable de igualación de los grupos. Para reducir el efecto producido por los ítems con funcionamiento diferencial, se han propuesto algunas técnicas de purificación que, en dos etapas o iterativamente, eliminan del criterio aquellos ítems que previamente han sido detectados con DIF (French y Maller, 2007, Gómez-Benito y Navas, 1996; Hidalgo y Gómez-Benito, 2003; Holland y Thayer, 1988; Navas-Ara y Gómez-Benito, 2002; Wang, Shih y Yang, 2009).

TIPOS DE DIF

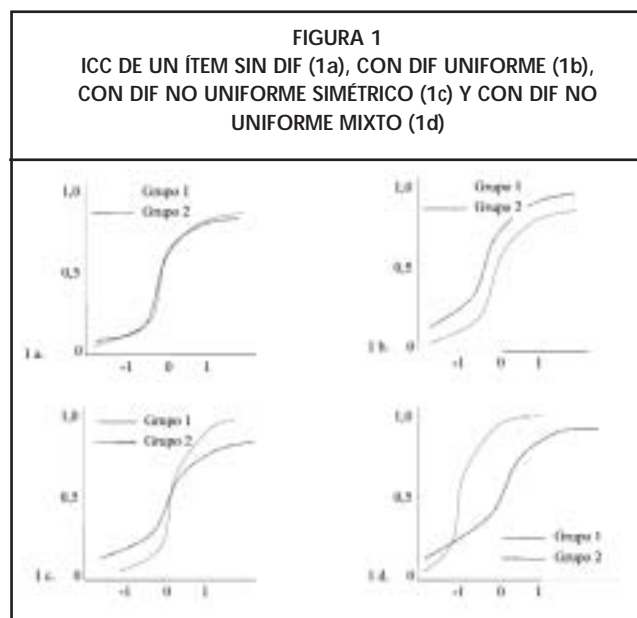
Aunque existen diversas taxonomías del DIF (ver Hessen, 2003), una clasificación muy extendida por su simplicidad proviene de Mellenbergh (1982). Este autor distingue dos tipos de DIF en función de la existencia o no de interacción entre el nivel en el atributo medido y el grupo de pertenencia de los individuos. En el denominado *uniforme* no existe interacción entre el nivel del rasgo medido y la pertenencia a un determinado grupo ya que la probabilidad de responder correctamente (o en una determinada dirección) al ítem es mayor para un grupo que para el otro de forma uniforme a lo largo de todos los niveles del rasgo. En el caso del DIF *no uniforme* sí que existe tal interacción, por lo que la probabilidad de cada grupo de responder correctamente (o en una determinada dirección) al ítem no es la misma a lo largo de los diferentes niveles del rasgo medido.

En el marco de la teoría de respuesta al ítem (véase Muñoz (2010) en este mismo número) se propone el concepto de *curva característica del ítem* (Item Characteristic Curve, ICC), de gran utilidad para entender gráficamente los diversos tipos de DIF. En ítems de respuesta dicotómica, la ICC relaciona la probabilidad de acertar el ítem (eje de ordenadas en el gráfico) con el nivel de los individuos en la variable medida o habilidad (eje de abscisas). De esta forma, un ítem no presenta DIF si su curva característica para el grupo focal y para el grupo de referencia coinciden (figura 1a), situación que se da cuando tanto el parámetro de dificultad (posición de la ICC en la escala de habilidad) como el de discriminación (proporcional a la pendiente de la ICC) presentan el mismo valor en ambos

grupos. El ítem muestra DIF *uniforme* si las respectivas ICCs no se cruzan en ningún nivel de la variable medida (figura 1b), hecho que se da cuando los parámetros de dificultad son distintos, pero los correspondientes parámetros de discriminación se mantienen iguales en ambos grupos. Finalmente, presenta DIF no uniforme si en algún punto las ICCs se cruzan. En este último caso, Swaminathan y Rogers (1990) establecen una segunda subdivisión. El DIF *no uniforme simétrico* quedaría representado por un cruzamiento central de las ICCs en el nivel de habilidad (figura 1c) y se da cuando el parámetro de dificultad se mantiene constante y el parámetro de discriminación varía entre los dos grupos, mientras que el DIF *no uniforme mixto* se da cuando los parámetros de dificultad y discriminación son distintos en los dos grupos y viene representado por un cruzamiento asimétrico de las ICCs del grupo focal y de referencia (figura 1d).

PROCEDIMIENTOS DE DETECCIÓN

Desde finales de los años 80 y durante toda la década de los 90, la elaboración y análisis de métodos y técnicas estadísticas para la detección y evaluación del DIF ha concentrado los esfuerzos de investigadores, y ha incrementado paulatinamente la sofisticación de los procedimientos utilizados. Su principal reto metodológico ha sido desarrollar procedimientos que, por un lado, sean sensibles a la detección tanto del DIF uniforme como no uniforme y, por otro lado, no confundan el DIF con el impacto. Además, como respuesta a la demanda progresiva de técnicas aplicables a ítems politómicos (como las



escalas tipo Likert), el interés se ha trasladado también al desarrollo de procedimientos útiles para este tipo de formato de respuesta, generalmente provenientes de extensiones de sus homólogos para ítems dicotómicos.

Teniendo en cuenta esta primera distinción sobre la naturaleza de respuesta al ítem (dicotómica/politómica), Potenza y Dorans (1995) clasifican los diferentes métodos en función del tipo de criterio de igualación de los grupos (puntuación observada/variable latente) y de la relación entre la puntuación en el ítem y la variable de igualación (paramétrica/no paramétrica). Basándose en esta taxonomía, Hidalgo y Gómez-Benito (2010) ofrecen una clasificación de todos los procedimientos actuales de detección del DIF.

En primer lugar, se puede estimar el nivel de habilidad de los individuos siguiendo dos estrategias: la primera, el *método de la variable latente* utiliza una estimación de la habilidad latente en el marco de la teoría de respuesta al ítem (TRI) mientras que el *método de la puntuación observada* consiste en utilizar la puntuación total observada del test. Un segundo criterio reside en cómo se estima la puntuación al ítem en cada uno de los niveles de habilidad. Una forma de proceder consiste en utilizar una función matemática que relacione la puntuación del ítem con el nivel de habilidad, como las ICCs de la figura 1, que representan gráficamente la probabilidad de obtener una determinada puntuación en el ítem en función del nivel de habilidad de los individuos. Como se ha comentado, diferencias en las ICCs de los grupos indican DIF y para que esto ocurra los parámetros que definen las correspondientes ICCs han de ser diferentes. En consecuencia, dado que las curvas vienen determinadas por uno o más parámetros en la función matemática, esta aproximación se denomina *método paramétrico*. En cambio, la segunda estrategia no utiliza ninguna función matemática para relacionar la respuesta al ítem con el nivel de habilidad, sino que simplemente tiene en consideración la puntuación observada al ítem en cada uno de los niveles de habilidad para cada grupo. En este caso, la presencia de DIF vendrá determinada por la obtención de diferencias entre grupos en la puntuación observada, sin tener en cuenta ningún modelo matemático (y, por tanto, tampoco parámetros). Por este motivo esta aproximación se conoce como el *método no paramétrico*. En tercer lugar se considera la naturaleza del tipo de respuesta, dicotómica o politómica. Dado que en el caso de ítems politómicos el DIF puede estar presente en las diferentes categorías de respuesta de un mismo ítem, y no necesariamente en la misma di-

rección ni en todas las categorías, las técnicas para ítems dicotómicos son siempre más sencillas computacional y conceptualmente que las extensiones para ítems de respuesta politómica.

Las técnicas que emplean la puntuación observada en el test como variable de igualación, asumiendo que esta puntuación es una estimación adecuada de la habilidad latente del individuo, pueden resultar imprecisas en la detección del DIF principalmente cuando el test contiene ítems de discriminación diversa, mientras que los métodos de la variable latente superan este inconveniente a base de incrementar la sofisticación de los modelos matemáticos de estimación de la habilidad. Una ventaja de los métodos no paramétricos, como el Mantel-Haenszel (MH) y el SIBTEST, es que los supuestos del modelo son escasos, por lo que el DIF no suele confundirse con la falta de ajuste del modelo. En el caso de los métodos paramétricos, como los procedimientos basados en la TRI, es necesario asegurar una adecuada estimación de los parámetros del test precisamente para evitar esta confusión, por tanto, se requieren tamaños muestrales del grupo de referencia y focal mucho más elevados que con los modelos no paramétricos.

Existe un abanico de programas informáticos que permite implementar la mayoría de procedimientos de detección del DIF. La mayor parte consisten en programas que han sido diseñados específicamente para la detección de los ítems con DIF, como MHDIF (Fidalgo, 1994), EZDIF (Waller, 1998a), DIFAS (Penfield, 2005) o EASYDIF (González, Padilla, Hidalgo, Gómez-Benito y Benítez, 2009) para el procedimiento MH y de libre distribución poniéndose en contacto con los autores del programa; DIF/DBF (Stout y Roussos, 1999) para el procedimiento SIBTEST, que se distribuye mediante Assessment System Corporation; RLDIF (Gómez-Benito, Hidalgo, Padilla, y González, 2005) para el procedimiento de Regresión Logística (RL), actualmente en proceso de comercialización; y IRTLDF (Thissen, 2001), TESTGRAPH (Ramsay, 2000) y LINKDIF (Waller, 1998b) para procedimientos basados en la TRI, también de libre distribución. Pueden utilizarse también recursos provenientes de programas estándares de análisis estadístico, que requieren licencia de uso, como por ejemplo SPSS (SPSS Inc., 2009) para MH y RL, LISREL (Jöreskog y Sörbom, 2006) o MPLUS (Muthén y Muthén, 1998, 2007) para procedimientos basados en modelos de ecuaciones estructurales.

Con la finalidad de estudiar su comportamiento, individual y comparativamente, numerosos trabajos se han

aproximado al estudio de las técnicas de detección del DIF mediante la simulación de datos, tanto en ítems dicotómicos como de respuesta politómica. Estos estudios básicamente analizan la variación en la tasa de falsos positivos o error Tipo I (detectar un ítem con DIF cuando en realidad no lo presenta) y de detecciones correctas o potencia estadística (identificar un ítem con DIF cuando realmente lo presenta) bajo diferentes condiciones de simulación, manipulando aquellas variables que supuestamente pueden modular las correspondientes tasas de detección (por ejemplo, el tamaño muestral, la contaminación del test o el tipo de DIF, entre otras) y observando los cambios producidos en ellas. Generalmente terminan el estudio delineando sugerencias y recomendaciones sobre las condiciones bajo las cuales el procedimiento en cuestión presenta un control de la tasa de error Tipo I y una adecuada potencia estadística. Una cuestión común a prácticamente la totalidad de estos estudios es que se centran en la detección del DIF en un único ítem. Hay que tener en cuenta que un test obviamente está compuesto por un conjunto de ítems, y que la dirección del DIF en los diversos ítems de un mismo test puede ser distinta (algunos pueden favorecer al grupo focal y otros al de referencia), de tal forma que los efectos individuales del DIF de los ítems se cancelen cuando se considera el test en global. Por tanto, en ocasiones, es interesante evaluar lo que se denomina el funcionamiento diferencial del test (Differential Test Functioning, DTF) o explorar el DIF en un subconjunto de ítems. En este contexto, algunas técnicas han procurado abordar específicamente el estudio del DIF en tests o conjuntos de ítems, como son el SIBTEST en ítems dicotómicos y el POLYSIBTEST en ítems politómicos, o la aproximación desde la TRI propuesta por Raju y su equipo de investigación (Oshima, Raju, y Nanda, 2006)

TAMAÑO DEL EFECTO

Otro tipo de estudios, también basados en la simulación de datos, aconsejan la inclusión de medidas del tamaño del efecto como complemento o alternativa a las pruebas de significación, a fin de poder evaluar la magnitud del efecto observado y comparar resultados obtenidos en diferentes estudios. Hay que tener en cuenta que detectar un ítem con DIF mediante una prueba de significación estadística no necesariamente implica que su efecto sea destacable, es decir, puede que su efecto sea de escasa relevancia. En este sentido, es importante examinar la magnitud del DIF porque los efectos de la presencia de ítems con DIF pueden ser triviales, cancelarse o pueden

realmente poner en duda las decisiones basadas en el test. La mayor parte de las técnicas de detección del DIF han propuesto diversas medidas. Por poner un ejemplo, Dorans y Holland (1993) presentan el estadístico Delta-DIF para el procedimiento Mantel-Haenszel, y dentro del ámbito de la regresión logística (RL) Zumbo y Thomas (1997) han sugerido el incremento en R^2 ; Gómez-Benito e Hidalgo (2007) y Monahan, McHorney, Stump y Perkins (2007) han propuesto el uso de la odds-ratio como medida del tamaño del efecto usando RL para ítems dicotómicos e Hidalgo, Gómez-Benito y Zumbo (2008) para ítems politómicos. Por norma general, estos trabajos establecen directrices o proponen criterios de clasificación que permiten interpretar los valores de la magnitud del DIF (no únicamente la presencia o ausencia de DIF) siguiendo las directrices de clasificación del *Educational Testing Service* que establece tres categorías, a saber: DIF insignificante (categoría A), DIF moderado (categoría B) y DIF elevado (categoría C). Los ítems que se clasifiquen como tipo C deben ser revisados y eliminados del test, por el contrario los ítems clasificados como tipo A y/o B pueden ser mantenidos en el test.

ELECCIÓN DE TÉCNICA

Aunque los avances en el desarrollo y la optimización de los métodos de detección han sido considerables, la idoneidad de aplicar un procedimiento determinado en una situación concreta todavía está llena de interrogantes. En este entramado de trabajos y técnicas, la duda suele trasladarse a la siguiente cuestión: ¿qué procedimientos empleamos con nuestros datos? La decisión de aplicar una técnica u otra suele basarse en diversos aspectos, dado que no existe hasta el momento ningún método que sea adecuado en la totalidad de situaciones. Se suelen tener en cuenta las diferencias en las distribuciones de habilidad de los grupos de referencia y focal, el tamaño muestral de ambos grupos, el tipo de DIF, la simplicidad computacional y disponibilidad de programas informáticos, y el criterio de igualdad de los grupos, entre otros. Y esta complejidad ha llevado a varios autores a pensar que la opción más conservadora consiste en aplicar diversas técnicas de detección del DIF y tomar la decisión última de mantener, reformular o eliminar el ítem en función de la convergencia o divergencia entre métodos de detección, teniendo en cuenta las características y peculiaridades de cada procedimiento. Parece evidente que si diversas técnicas coinciden en sus decisiones, se tiene más certeza de la presencia o ausencia de DIF, mientras que si existe divergencia entre técnicas

deberíamos fijarnos en las características de los procedimientos de detección empleados. En cualquier caso se trataría de acumular evidencias en una dirección u otra, como en todo procedimiento de validación de un instrumento.

Siguiendo la clasificación de métodos de detección basada en el tipo de criterio de igualación de los grupos (puntuación observada/variable latente) y la relación entre la puntuación en el ítem y la variable de igualación (paramétrica/no paramétrica), se ha visto que se establecen cuatro tipos de métodos de detección tanto para ítems dicotómicos como de respuesta politómica: i) puntuación observada/paramétrica, ii) puntuación observada/no paramétrica, iii) variable latente/paramétrica, y iv) variable latente/no paramétrica. Ya se ha señalado más arriba que los métodos que emplean la puntuación observada como estimación de la habilidad de los sujetos pueden resultar imprecisos cuando el criterio de igualación presenta un porcentaje elevado de ítems que funcionan diferencialmente, mientras que los métodos de la variable latente pueden superar este inconveniente a base de incrementar la complejidad matemática. Pero una ventaja de los métodos no paramétricos es que los supuestos del modelo son escasos, por lo que el DIF no suele confundirse con la falta de ajuste del modelo, mientras que con los métodos paramétricos es necesario asegurar un adecuado ajuste del modelo para evitar esta confusión, por tanto, se requieren tamaños muestrales mucho más elevados que con los modelos no paramétricos. Teniendo en cuenta las ventajas e inconvenientes generales que implican los distintos tipos de técnicas de detección, una recomendación iría en la línea de tomar la decisión última en base a la aplicación de una técnica de cada uno de los cuatro tipos existentes, por ejemplo, una opción sería emplear en la detección de ítems dicotómicos RL, MH, TRI y SIBTEST. Sin embargo, habría que tener en cuenta otras consideraciones respecto a los datos que podrían proporcionar una explicación sobre las posibles divergencias entre métodos de detección.

La primera de ellas deriva del tamaño muestral. Si se trabaja con tamaños reducidos, se ha evidenciado que RL y MH funcionan adecuadamente para ítems dicotómicos (Muñiz, Hambleton, y Xing, 2001; Swaminathan y Rogers, 1990) y TRI (usando la prueba de razón de verosimilitud) para ítems politómicos (Bolt, 2002). Si se dispone de tamaños considerables se puede optar por otras técnicas, como el SIBTEST y el POLYSIBTEST para la detección de ítems dicotómicos y politómicos.

Otro consejo giraría entorno al tipo de DIF. Existen técnicas que han sido diseñadas específicamente para la detección del DIF uniforme, por lo que pueden presentar ciertas dificultades en la detección del DIF no uniforme, mientras que otras se han propuesto para la detección de ambos tipos de DIF. Cuando se intuye la presencia de DIF no uniforme, es preferible emplear técnicas que sean sensibles a este tipo de funcionamiento diferencial. De nuevo, con tamaños muestrales reducidos, se puede optar por RL en ítems dicotómicos (Hidalgo y López-Pina, 2004) y TRI (usando la prueba de razón de verosimilitud) en ítems politómicos (Bolt, 2002). Si se emplean tamaños considerables se pueden seleccionar otras técnicas como el SIBTEST en el caso de ítems dicotómicos y la regresión logística multinomial (Zumbo, 1999) o el DFIT (Oshima, Raju y Nanda, 2006) para los de respuesta politómica.

Por otro lado, constatamos que la mayoría de los métodos actuales para detectar DIF requieren que el test a analizar contenga un número elevado de ítems (p.e. mayor de 30) para que el resultado sea fiable. Por el contrario, los cuestionarios y encuestas que se suelen utilizar en el ámbito de las ciencias sociales y de la salud suelen tener un número pequeño de ítems (entre 5 y 30 ítems). Cuando trabajamos con tests tan cortos la fiabilidad de las puntuaciones es menor y por lo tanto los errores de medida mayores. Métodos tales como RL o MH, que utilizan la puntuación observada en el test como variable de equiparación en el análisis del DIF, pueden ver seriamente afectada su eficacia para detectarlo. El uso de los modelos MIMIC (Gelin y Zumbo, 2007) es una alternativa.

Dado que la mayoría de estudios postulan que con la aplicación de procedimientos de purificación del criterio se produce una reducción de la tasa de falsos positivos y un incremento de la potencia estadística de diversos métodos, es aconsejable su empleo. Finalmente, en la medida de lo posible, se recomienda acompañar las tasas de detección con alguna medida del tamaño del efecto.

TESTS JUSTOS

Ya se ha comentado cómo en la década de los sesenta en EEUU se empezó a cuestionar el uso de los tests para evaluar de modo equitativo a distintos grupos de sujetos y cómo el artículo de Jensen (1969) sobre la naturaleza hereditaria de la inteligencia agudizó la polémica entre ambientalistas y genetistas. Dicha polémica tuvo una relevante repercusión social y política, llegándose a considerar que los tests en los que se constataban diferencias en función de características socioeconómicas o raciales,

estaban sesgados y eran injustos. Esta repercusión llegó a los tribunales donde se fallaron sentencias en contra de decisiones de selección de personal o de admisión a instituciones educativas. Una de las consecuencias más relevantes fue la llamada "regla dorada" que surgió del acuerdo el Educational Testing Service (la compañía de tests más importante de EEUU) y la compañía de seguros Golden Rule, por la que se debían eliminar los ítems en los que los sujetos de raza blanca obtuvieran un resultado superior en un 15% a los de raza negra. Evidentemente, dicha regla, basada únicamente en el índice de dificultad de los ítems para distintos grupos, podía conllevar la eliminación de ítems con alto poder discriminativo respecto al rasgo medido.

En ese momento, los términos sesgo e injusticia se equipararon y no se disponía de criterios eficaces para identificar si el comportamiento diferencial del test era debido a diferencias reales en el rasgo o a diferencias artefactuales provocadas por el instrumento utilizado. Esta oposición a que los tests se utilizaran para tomar decisiones que afectarían a la vida laboral o vida académica, fue también acicate para que los psicómetras se esforzaran en ofrecer definiciones y técnicas de detección de sesgos, dando lugar a una de las líneas de investigación psicométrica más fructífera de las últimas décadas. Así, en los años setenta-ochenta, aparece el término "funcionamiento diferencial del ítem" y se le distingue del término "sesgo", se ponen de relieve las diferencias entre DIF e impacto y se proponen técnicas de detección que permitan deslindar ambos aspectos. En la década de los noventa se incide en la explicación del DIF mediante la dimensionalidad de los tests; así, Ackerman (1992) distingue entre habilidad objetivo (aquella que pretende medir el test) y habilidad ruido (que no se pretendía medir pero que puede influir en las respuestas a algunos ítems del test): el DIF se puede presentar si los ítems del test miden una habilidad ruido en la que los sujetos difieren en función del grupo. Roussos y Stout (1996) añaden un matiz más: cambian la terminología y hablan de habilidades secundarias en vez de habilidad ruido, distinguiendo entre DIF-benigno y DIF-adverso; se da DIF-benigno cuando la habilidad secundaria es una dimensión auxiliar que se pretendía medir y DIF-adverso cuando la habilidad secundaria es una habilidad ruido.

De todos modos, y si bien es crucial ofrecer procedimientos estadísticos capaces de una eficaz detección de los ítems con funcionamiento diferencial, éstos por sí mismos no ofrecen una explicación de por qué el DIF se

produce y si implica o no un sesgo. No hay que olvidar que la presencia de DIF es una condición necesaria pero no suficiente para poder hablar de sesgo del ítem: el DIF existe cuando individuos con una habilidad comparable pero de grupos distintos responden diferencialmente al ítem, mientras que para que exista sesgo se requiere además que estas diferencias sean atribuibles a alguna característica del ítem ajena al atributo medido por el test.

A finales del siglo pasado y principios de éste, se ha incidido en la importancia de analizar las causas del DIF. En el contexto de la adaptación de tests, los estudios de Allalouf, Hambleton y Sireci (1999) y de Gierl y Khaliq (2001) aportan algunas posibles causas centradas en el formato y el contenido del ítem; Zumbo y Gelin (2005) recomiendan que se consideren además diversas variables contextuales. Sin embargo, Ferne y Rupp (2007), en una revisión de 27 estudios que intentan identificar causas de DIF, constatan que los avances logrados son poco relevantes. Este es quizás uno de los retos actuales de la investigación en DIF, que merecería el mismo ahínco investigador que los anteriores problemas mencionados que se han ido solventando. Para ello convendría llevar a cabo estudios expresamente diseñados para investigar las causas del DIF, y sin duda la teoría multidimensional podría orientar la búsqueda de las causas hacia las habilidades espúreas que se distribuyen de forma distinta entre los grupos comparados. Considerar un resultado de DIF como una evidencia de sesgo implica explicar por qué el rasgo es multidimensional para un subgrupo específico y elaborar un argumento razonando la irrelevancia de la fuente de DIF para este rasgo (Camilli y Shepard, 1994).

En última instancia, como cualquier otro aspecto de la validez, el análisis del DIF es un proceso de acumulación de evidencias. Valorar e interpretar dichas evidencias requiere del juicio racional de los expertos y no existe una única respuesta correcta. En este sentido hay que apostar por la responsabilidad profesional de las personas que emplean tests, sensibilizarse y formarse acerca de la relevancia de la calidad métrica de estos instrumentos de medida, con la finalidad última de garantizar un proceso adecuado y justo de medición. Como paso previo a la aplicación de los métodos de detección, Hambleton y Rogers (1995) desarrollaron una lista de indicadores que pueden hacer sospechar de la posible presencia de DIF, por ejemplo, ítems que asocian los hombres al deporte y las mujeres a las acti-

vidades de la casa, o que utilizan ciertas palabras cuyo significado es más familiar para una cultura que para otra (alimentos, juegos, enfermedades, eventos históricos, etc.), entre otros. Además, Hambleton (2006) recomienda que tanto los creadores como los usuarios de tests tengan en cuenta los estudios previos de DIF, ya que pueden proporcionar información acerca de las características comunes a los ítems con DIF, así como las particularidades que comparten los ítems sin funcionamiento diferencial. Esta información es crucial tanto para el desarrollo de nuevos ítems como para alertarnos de la posible presencia de DIF en las pruebas existentes.

Como señala Zumbo (2007), los métodos de detección del DIF y del sesgo de los ítems se utilizan típicamente en el proceso de análisis de los ítems cuando se desarrollan nuevos instrumentos de medida, se adaptan tests existentes a un nuevo contexto de evaluación o a otras poblaciones que no se tuvieron en cuenta en el momento de crear el instrumento, se adaptan pruebas ya existentes a otras lenguas o culturas, o se validan las inferencias derivadas de las puntuaciones del test. Se constata pues que el ámbito de aplicación de un análisis del DIF es extenso y que está presente en las distintas fases de creación y adaptación de un instrumento de medida. En España, donde mayoritariamente se importan tests, es especialmente importante el análisis del DIF y del sesgo en la adaptación de instrumentos estandarizados a la lengua y contexto cultural propios. Desde el Colegio Oficial de Psicólogos (COP) se ha participado en la creación de unas directrices precisamente dedicadas a la creación y adaptación de tests, y obviamente en ellas el DIF tiene un papel destacado (Muñiz y Hambleton, 1996).

También tiene un papel relevante en los últimos estándares (APA et al, 1999) que incluyen el análisis del DIF y del sesgo en el análisis de la validez, concretamente en las evidencias basadas en la estructura interna del test. En definitiva, la decisión sobre si el resultado obtenido en un estudio es o no evidencia de sesgo sólo se puede tomar desde la teoría de la validez: conociendo la teoría subyacente al test, la interpretación que se pretende hacer de las puntuaciones y el contexto en el que se utiliza el test; en este sentido la ampliación de los contenidos de la validez permite que los estudios de sesgo aborden la perspectiva social del problema como una faceta más del proceso de validación de un test. El artículo de Prieto y Delgado (2010) en este mismo número describe el proceso de validación con más detalle.

PARA SABER MÁS

Aquel lector interesado en profundizar en el conocimiento de las técnicas de detección del DIF así como las implicaciones prácticas que supone la presencia de DIF en los ítems de un test puede consultar diversas revisiones teóricas (Camilli y Shepard, 1994; Fidalgo, 1996; Gómez-Benito e Hidalgo, 1997; Hidalgo y Gómez-Benito, 1999, 2010; Osterlind y Everson, 2009; Penfield y Lam, 2000; Potenza y Dorans, 1995) que se aproximan al estudio de las distintas técnicas de forma narrativa, exponiendo los procedimientos, especificando las ventajas y desventajas de su aplicación, y delineando recomendaciones para su empleo.

AGRADECIMIENTOS

Este trabajo ha sido financiado, en parte, por el Ministerio de Ciencia e Innovación (PSI2009-07280) y, en parte, por la Generalitat de Catalunya (2009SGR00822). Los autores muestran asimismo su agradecimiento a Vicente Ponsoda por su invitación a participar en este monográfico y al resto de investigadores de esta edición por su colaboración y ayuda en que este proyecto haya salido adelante.

REFERENCIAS

- Ackerman, T.A. (1992). A didactic explanation of items bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.
- Allalouf, A., Hambleton, R. K. y Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.
- American Psychological Association, American Educational Research Association y National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141.
- Camilli, G. y Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.
- Dorans, N. J., y Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

- Ferne, T. y Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4, 113-148.
- Fidalgo, A.M. (1994). MHDIF – A computer-program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement*, 18(3), 300-300.
- Fidalgo, A. M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), *Psicometría* (pp. 370-455), Madrid: Universitas.
- French, B.F. y Maller, S.J. (2007). Iterative purification and effect size use with Logistic Regression for Differential Item Functioning Detection. *Educational and Psychological Measurement*, 67, 373-393.
- Gelín, M.N. y Zumbo, B.D. (2007). Operating characteristics of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation for short scales. *Journal of Modern Applied Statistical Methods*, 6, 573-588.
- Gierl, M. J., y Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement*, 38, 164-187.
- Gómez-Benito, J., e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología*, 74(3), 3-32.
- Gómez-Benito, J. e Hidalgo, M.D. (2007). Comparación de varios índices del tamaño del efecto en regresión logística: Una aplicación en la detección del DIF. Comunicación presentada en el X Congreso de Metodología de las Ciencias Sociales y de la Salud, Barcelona, 6-9 febrero.
- Gómez-Benito, J., Hidalgo, M. D., Padilla, J. L., y González, A. (2005). Desarrollo informático para la utilización de la regresión logística como técnica de detección del DIF. Demostración informática presentada al IX Congreso de Metodología de las Ciencias Sociales y de la Salud, Granada, España.
- Gómez-Benito, J., y Navas, M.J. (1996). Detección del funcionamiento diferencial del ítem: Purificación paso a paso de la habilidad. *Psicológica*, 17, 397-411.
- González, A., Padilla, J.L, Hidalgo, M.D., Gómez-Benito, J. y Benítez, I. (2009) EASY-DIF: Software for analysing differential item functioning using the Mantel-Haenszel and standardization procedures. *Applied Psychological Measurement*. (Enviado para su publicación).
- Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44(11 Suppl. 3), S182-S188.
- Hambleton, R.K., y Rogers, H.J. (1995). Item bias review (EDO-TM-95-9). Washington, DC: Clearinghouse on Assessment and Evaluation.
- Hessen, D.J. (2003). *Differential item functioning: Types of DIF and observed score based detection methods*. Dissertation (supervisors: G.J. Mellenbergh & K. Sijtsma). Amsterdam: University of Amsterdam.
- Hidalgo, M. D., y Gómez-Benito, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politémicos. *Metodología de las Ciencias del Comportamiento*, 1(1), 39-60.
- Hidalgo, M. D., y Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment*, 19(1), 1-11.
- Hidalgo, M. D., y Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd edition). USA: Elsevier - Science & Technology.
- Hidalgo, M.D., Gómez-Benito, J. y Zumbo, B.D. (2008). Efficacy of R-square and Odds-Ratio effect size using Discriminant Logistic Regression for detecting DIF in polytomous items. Paper presented at the 6th Conference of the International Test Commission, 14-16 July, Liverpool, UK.
- Hidalgo, M. D., y López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, 64(4), 903-915.
- Holland, P., y Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H. I. Braun (Eds.), *Test Validity* (pp.129-145). Hillsdale, NJ: LEA.
- Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 39(1), 1-123.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Jöreskog, K.G., y Sörbom, D. (2006). *Lisrel 8 (version 8.8)*. Chicago, Illinois: Scientific Software International, Inc.
- Mellenbergh, G. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105-118.
- Messick, S. (1989). Validity. En R. Linn (Ed.). *Educatio-*

- nal measurement (3rd edition, pp. 13-104). Washington, DC: American Council on Education.
- Monahan, P.O., McHorney, C.A., Stump, T.E. y Perkins, A.J. (2007). Odds-ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Behavioral Statistics*, 32, 1, 92-109.
- Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo*, 31(1), 57-66.
- Muñiz, J., y Hambleton, R.K. (1996). Directrices para la traducción y adaptación de los tests. *Papeles del Psicólogo*, 66.
- Muñiz, J., Hambleton, R. K., y Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, 1(2), 115-135.
- Muthén, L.K., y Muthén, B.O. (1998, 2007). *MPLUS statistical analysis with latent variables. User's Guide*. Los Angeles, CA: Muthén and Muthén.
- Navas-Ara, M. J. y Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment*, 18(1), 9-15.
- Oshima, T. C, Raju, N. S. y Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of item and tests (DFIT) framework. *Journal of Educational Measurement*, 43, 1-17.
- Osterlind, S. J., y Everson, H. T. (2009). *Differential item functioning* (2nd edition). Thousand Oaks, California: Sage Publications, Inc.
- Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement*, 29(2), 150-151.
- Penfield, R. D., y Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, 19, 5-15.
- Potenza, M., y Dorans, N. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19, 23-37.
- Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, 31(1), 67-74.
- Ramsay, J. O. (2000). *TestGraph: A program for the graphical analysis of multiple choice and test questionnaire*. Unpublished manual.
- Roussos, L. y Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, 20, 355-371.
- SPSS 15.0. (2009). SPSS Inc. 1989-2009.
- Stout, W. y Roussos, L. (1999). *Dimensionality-based DIF/DBF package* [Computer Program]. William Stout Institute for Measurement. University of Illinois.
- Swaminathan, H. y Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.
- Thissen, D. (2001). *IRTLRDIF v2.0b. Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio Test for Differential Item Functioning*. Available on Dave Thissen's web page.
- van de Vijver, F., y Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London: Sage Publications.
- Waller, N. G. (1998a). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and Logistic Regression procedures. *Applied Psychological Measurement*, 22, 391.
- Waller, N.G. (1998b). LINKDIF: Linking item parameters and calculating IRT measures of Differential Item Functioning of Items and Tests. *Applied Psychological Measurement*, 22, 392.
- Wang, W.-C, Shih, C.-L. y Yang, C.-C. (2009). The MI-MIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, 69, 713-731.
- Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223-233.
- Zumbo, B. D. y Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5, 1-23.
- Zumbo, B. D., y Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.