Psicothema
2002. Vol. 14, nº 2, pp. 491-496

ISSN 0214 - 9915 CODEN PSOTEG

Supuestos y consideraciones en los estudios empíricos sobre el funcionamiento diferencial de los items

Ángel M. Fidalgo Aliste y Doris Ferreres Traver*
Universidad de Oviedo y * Universidad de Valencia

En este artículo se analizan tres de los supuestos más comunes en los estudios sobre el funcionamiento diferencial de los ítems (DIF), y las consecuencias de su incumplimiento. Primero, que las muestras utilizadas para evaluar el DIF son muestras estadísticamente representativas de las poblaciones de interés. Segundo, que la mayor parte de los ítems que componen el test son ítems válidos. Tercero, que el coste de cometer un error de Tipo I es mayor que el coste de cometer un error de Tipo II. También se analizan las implicaciones que la significación estadística y práctica, las características de los ítems y de las poblaciones, y, finalmente, la concordancia entre los procedimientos de detección, tienen en los estudios empíricos sobre DIF. En resumen, se indican algunas precauciones que hay que tener presentes al analizar estadísticamente el DIF en datos empíricos.

Assumptions and considerations for detecting differential item functioning in empirical data. This paper examined three of the most common assumptions in DIF analysis, and the consequences of their violation. First, the samples used for DIF analysis are statistically representative samples of the specified populations. Second, the vast majority of items in the test are valid items. Third, the cost of making a type I error is greater than the cost of making a type II error. As well, the implications that for empirical DIF studies show the statistical and practical significance, the characteristics of items and populations, and finally, the coherence between complementary tecniques, were examined. In short, some cautions are offered regarding the use of statistical DIF analysis with empirical data.

La estructura del mundo quizá sea entera y perfecta como una bola de metal pulido. Para los que formamos parte de ella se revela, sin embargo, confusa y cambiante. Heráclito expresó el principio ontológico del dominio del cambio en el mundo, de una forma muy bella, empleando la discordia o la guerra como una metáfora del cambio: πόλεμος πάντων μὲν πατήρ ἐστι, πάντων δὲ βασιλεύς, καὶ τοὺζ μὲν θεοὺς ἔδειξε τοὺζ δὲ ἀνθρώπους, τοὺς μὲν δούλους ἐποίησε τοὺς δὲ ἐλευθέρους (La guerra es el padre y el rey de todas las cosas; a unos los muestra como dioses y a otros como hombres, a unos los hace esclavos y a otros libres). Para aquellos racionalistas a los que un presocrático no ofrezca las suficientes garantías, podemos aducir como prueba de la naturaleza cambiante del mundo el que la ciencia, aunque guiada por una voluntad de lo verdadero, sólo acierte a plantear hipótesis, y que gran parte de sus explicaciones se hagan mediante modelos que incorporan distinto grado de incertidumbre, esto es, mediante modelos probabilísticos. Se preguntarán, ¿y lo dicho a cuenta de qué? Quizá pueda omitirse a Heráclito, quizá sobre hablar del mundo. Es posible que para sacar agua de un pozo sólo se necesite un cántaro y unos buenos brazos. A veces conviene, sin embargo, describir el paisaje, situar el pozo…, no nos perdamos con el cántaro camino de la fuente. En todo caso, la diferencia entre lo que hay, que es variable y desconocido (y lo dicho viene a cuenta de ello) y lo que uno construye, simplificado y ordenado de antemano según ciertos fines, nos servirá para caracterizar dos formas de aproximarse al estudio del funcionamiento diferencial de los ítems (DIF) bien distintas: los estudios de simulación y los estudios empíricos.

Los estudios de simulación pueden considerarse diseños experimentales en los que el investigador intenta determinar el efecto que las variables independientes (VIs) por él manipuladas (método de detección del DIF, tamaño de muestra, magnitud de DIF, etc.) tienen sobre las variables dependientes (VDs) de interés. Las VDs más frecuentemente consideradas han sido la capacidad de los diversos métodos para detectar el DIF correctamente (potencia de prueba) y el grado en que realizan detecciones incorrectas (la tasa de error de Tipo I). Este tipo de investigaciones permiten establecer con garantías las relaciones encontradas entre las VIs y las VDs, con la contrapartida de que dicha seguridad deberá circunscribirse a los niveles de las VIs manipulados, ya que éstos suelen ser elegidos arbitrariamente. Así, podremos determinar con certeza qué técnicas son mejores que otras y en qué condiciones, y esto porque conocemos la verdadera naturaleza (los parámetros que determinan la simulación) de los datos analizados.

Por el contrario, en los estudios empíricos, a partir de datos ajenos a nuestra voluntad y conocimiento, debemos determinar qué ítems presentan DIF, y esto sin cometer, a ser posible, ninguno de estos dos errores: detectar incorrectamente a ítems sin DIF (error

Fecha recepción: 21-6-01 • Fecha aceptación: 23-10-01
Correspondencia: Ángel M. Fidalgo Aliste
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
E-mail: fidalgo@correo.uniovi.es

de Tipo I), o que dejemos sin detectar a ítems con DIF (error de Tipo II). Como sabemos que el objetivo es imposible, donde dice 'sin cometer', diremos, cometiendo dichos errores en la menor cuantía posible. Y ya es mucho y muy difícil, y pese a ello, dichas investigaciones se suelen acometer de forma bastante estandarizada mediante la aplicación de alguna prueba de significación estadística utilizando un nivel de significación del 0.05 o del 0.01, las más de las veces. Esta forma de evaluación del DIF un tanto mecánica, conlleva la asunción de ciertos supuestos que a veces pueden no ser muy convenientes dadas las características de los datos que se quieren analizar. A continuación se explicitarán tres de los supuestos más comunes en las investigaciones sobre el DIF, y las consecuencias de su incumplimiento.

Primer supuesto

Las muestras utilizadas para evaluar el DIF son muestras estadísticamente representativas de las poblaciones de interés.

Podemos decir que un ítem funciona diferencialmente cuando examinados de igual nivel en la variable o variables medidas por el test, pero pertenecientes a diferentes *poblaciones*, no tienen la misma probabilidad de resolverlo correctamente. Nótese como en la definición se hace referencia al funcionamiento diferencial del ítem en la población, que es lo que realmente nos interesa conocer y para ello aplicamos pruebas de significación estadística, y no en la muestra, donde ya conocemos con certeza como se comporta el ítem. Por lo tanto, que las inferencias realizadas sobre las poblaciones de interés (varones frente a mujeres, castellanoparlantes frente a valencianoparlantes, etc.) sean o no correctas dependerán, en primera instancia, de que el muestreo se haya realizado correctamente. Generalizar a las poblaciones a partir de muestras no representativas es similar a atribuir riqueza o pobreza a un hombre a partir del saldo de la cuenta corriente de su vecino. Todo empieza, y se supone, por un buen muestreo.

Segundo supuesto

La mayor parte de los ítems que componen el test son ítems válidos.

Resumiendo un poco el procedimiento, todas las técnicas de detección del DIF utilizan la misma estrategia: comparar el desempeño en el ítem de grupos de sujetos pertenecientes a distintas poblaciones, pero con el mismo nivel en la variable que presumiblemente mide el test. Como estimación del nivel de la variable medida por el test en la mayor parte de los estudios se utiliza, bien la puntuación total en el test, bien la estimación de la variable latente (θ) dentro de la Teoría de Respuesta al Ítem (TRI) . Es decir, se analizan los ítems sin relación a ningún criterio externo al test evaluado. Por ello, para garantizar que disponemos de medidas insesgadas del nivel de competencia de los examinados, es necesario asumir que la mayor parte de los ítems que componen el test son válidos. El incumplimiento de este supuesto conlleva una disminución de la potencia de prueba y, sobre todo, un incremento en la tasa de error de Tipo I (Fidalgo, Mellenbergh y Muñiz, 1998, 1999, 2000). Dado que es posible encontrar en tests de rendimiento porcentajes de ítems sesgados de hasta el 20% y el 40% debido a diferencias instruccionales (Miller y Linn, 1988), diversos autores han propuesto aplicar los métodos para detectar el DIF de forma iterativa. Esto significa realizar análisis sucesivos hasta llegar a identificar un conjunto de ítems sin DIF que puedan ser utilizados como criterio para evaluar el funcionamiento diferencial del resto de los ítems del test. Ya que en los estudios empíricos es imposible determinar a priori el porcentaje de ítems del test que pueden funcionar diferencialmente, y se anda como por camino desconocido en noche cerrada, la utilización de procedimientos iterativos es altamente recomendable.

Dice el encabezamiento de este apartado: 'La mayor parte de los ítems que componen el test son ítems válidos'. Hasta el momento nos hemos centrado en las implicaciones que se derivan de la expresión 'la mayor parte'. Ahora analizaremos el adjetivo 'válidos'. Se dice que un test o un ítem es válido cuando mide el constructo que pretende medir. No dice nada la definición sobre si lo que mide es uno o vario. La mayor parte de los estudios empíricos asumen, sin embargo, que sólo se mide un constructo, es decir, asumen la unidimensionalidad del test. Cuando aplicamos para detectar el DIF el procedimiento Mantel-Haenszel (MH) o el SIBTEST, por poner dos ejemplos, y lo hacemos utilizando como variable para equiparar a los examinados la puntuación total en el test, estamos asumiendo que el constructo medido por el test es básicamente unidimensional. En estas condiciones, y tal y como establece la teoría multidimensional del DIF (Akerman, 1992; Kok, 1988; Shealy y Stout, 1993; Roussos y Stout, 1996a), si ciertos ítems del test miden otras variables no deseadas (variables espurias) además de la variable principal que pretende medir el test, entonces esos ítems pueden llegar a presentar DIF (para un análisis más detallado véase Fidalgo, 1996, pág. 381). Existe, sin embargo, la posibilidad de que el test esté diseñado para medir más de una variable, es decir, que el test sea *intencionadamente multidimensional*. Por ejemplo, es perfectamente factible que un test de matemáticas compuesto por ítems de geometría, álgebra y trigonometría sea un test multidimensional. En este caso, el test estaría midiendo, o tres dimensiones principales (geometría, álgebra y trigonometría), o una dimensión principal de matemáticas compuesta de tres dimensiones auxiliares que son pertinentes. En estas condiciones, si realizásemos un análisis de DIF empleando la puntuación total en el test como variable de bloqueo, es muy posible que encontráramos que los ítems de geometría presentan DIF contra las mujeres (ya que los hombres obtienen por término medio puntuaciones mayores en geometría que las mujeres) y los de álgebra presentan DIF contra los hombres (ya que las mujeres obtienen por término medio mayores puntuaciones en álgebra que los hombres; O'Neill y McPeeks, 1993). Si por el contrario, como debería hacerse, utilizásemos la puntuación en cada una de esas dimensiones para analizar los ítems correspondientes, es decir, realizásemos los análisis sobre conjuntos de ítems dimensionalmente homogéneos, veríamos desaparecer el DIF.

Por consiguiente, ni se debe confundir la multidimensionalidad con el DIF, ni se pueden intercambiar estos términos como sinónimos. Una excelente guía teórica sobre este particular puede encontrase en Roussos y Stout (1996a) y Gierl, Bisanz, Bisanz, Boughton, y Khaliq (2001), y un excelente ejemplo práctico en Ryan y Chiu (2001). Señalar, por último, que en la literatura existen técnicas expresamente formuladas para evaluar el DIF en situaciones multidimensionales (Flowers, Oshima y Raju, 1999; Stout, Li, Nandakumar & Bolt, 1997), o ejemplos de aplicación de otras tan comunes como el procedimiento MH o la regresión logística (Mazor, Kanjee & Clauser, 1995; Mazor, Hambleton y Clauser, 1998). Como corolario, concluiremos que en los estudios empíricos un paso previo a cualquier análisis de DIF deberá ser siempre el de determinar la estructura dimensional del test; y a ser

posible, combinando los análisis teóricos y de contenido con los análisis estadísticos pertinentes (para una revisión de los procedimientos estadísticos para determinar la dimensionalidad de los datos, véase Hattie,1985, y Cuesta, 1996).

Tercer supuesto

El coste de cometer un error de Tipo I (detectar a los ítems que no tienen DIF) es mayor que el coste de cometer un error de Tipo II (no detectar a los ítems con DIF).

Como sabemos, la probabilidad de cometer un error de Tipo II (β) y la probabilidad de rechazar correctamente la hipótesis nula (H_0) o potencia de prueba ($1-\beta$) son función: a) del nivel de significación adoptado (α); b) del tamaño de muestra (N); c) de la magnitud de la desviación típica poblacional del estadístico de contraste (σ); d) de la diferencia entre el verdadero valor del parámetro y el valor hipotetizado $|\theta - \theta_0|$, es decir, del tamaño del efecto; y e) de la decisión de probar una hipótesis unidireccional o bidireccional (la probabilidad de error de Tipo II es mayor en este último caso). Sabemos que los factores α, β , $1-\beta$, σ, N y $|\theta - \theta_0|$ están interrelacionados. Si conocemos cualesquiera cuatro podríamos obtener el quinto. ¿A cuál prestar más atención? Tradicionalmente los procedimientos estadísticos se han preocupado de minimizar los errores de Tipo I, prestando mucha menos atención a los errores de Tipo II, y ésta es la tónica que siguen los estudios de DIF. Esta especial atención al nivel de significación, se ha visto favorecida, sin duda, por dos componentes bien arraigados en la naturaleza humana: la comodidad (α se fija a priori sin necesidad de ningún cálculo) y el gregarismo (se suelen utilizar por convención dos niveles, el 0.05 y el 0.01). Una convención similar afecta al error de Tipo II, para el que suelen considerarse adecuados valores iguales o menores a 0.2, lo que nos da una potencia de prueba de 0.8. Estas convenciones, como todas las demás, desde la utilización del chaqué en las bodas al establecimiento del kilómetro cero en Madrid, tienen sus consecuencias. En este caso implican que el investigador está penalizando la posibilidad de cometer un error de Tipo I frente un error de Tipo II. Así por ejemplo, si en los análisis estadísticos de una investigación usan un nivel de significación de 0.05 y β= 0.2, eso significa que consideran, aunque quizá no lo sepan, que es cuatro veces peor (0.2/0.05) cometer un error de Tipo I (rechazar la hipótesis H_0 cuando es verdadera) que uno de Tipo II (aceptar la H_0 cuando es falsa). Es decir, que es cuatro veces peor eliminar erróneamente a un ítem que no presenta DIF, que el que se nos pase por alto un ítem que funcione diferencialmente. Esta razón de costes quizá sea pertinente y buen reflejo del elevado precio que pueden suponer los falsos positivos cuando el DIF se evalúa en una prueba que ya está baremada y en circulación, pero es a todas luces excesiva cuando se está evaluando el DIF en las etapas iniciales de construcción de un test, donde el coste de sustitución de los ítems es mucho menor. En esa última situación utilizar niveles de significación más elevados puede estar sobradamente justificado. Además, téngase presente que la razón de coste entre ambos errores se incrementa conforme la potencia de prueba disminuye (véase la Tabla 1), y que en muchas situaciones ésta es más bien baja (Fidalgo, Mellenbergh y Muñiz, 1999).

Tabla 1
Razón de coste (β/α) entre el error de Tipo II y el de Tipo I, para distintas potencias de prueba ($1-\beta$) y niveles de significación (α)

	α		
1-β	**0.01**	**0.05**	**0.10**
0.8	20	4	2
0.6	40	8	4
0.4	60	12	6

Junto con los aspectos hasta ahora indicados, a la hora de llevar a cabo una investigación empírica deberíamos tener presentes, además, las siguientes consideraciones:

Primera consideración

Significación estadística y significación práctica

Un error muy extendido, al menos entre los estudiantes de licenciatura, es interpretar la probabilidad de ocurrencia de la H_0 como un índice de la magnitud de los tratamientos, y concluir, por ejemplo, que éstos son muy eficaces cuando en la salida de los programas estadísticos se observan valores de p= 0.000. Llevado a nuestro caso, dirían que cuanto menor sea dicha probabilidad mayor DIF presentará el ítem. Ya hemos señalado que la significación estadística está determinada por más factores que el tamaño del efecto y, por lo tanto, que esa interpretación es incorrecta. Una cosa es la significación estadística que establece cuál es la probabilidad de obtener dichos resultados siendo la H_0 verdadera, y otra la significación práctica que indica si el efecto observado es suficientemente grande para tener alguna utilidad. Afortunadamente todos los manuales que tratan sobre el DIF recalcan la importancia que tiene el cálculo del tamaño del efecto (la magnitud del DIF) en este tipo de investigaciones. Camilli y Shepard (1994), por ejemplo, señalan que sería mucho más correcto decir que las medidas de DIF deben complementarse con tests estadísticos que lo inverso. Nosotros, huelga decirlo, pensamos lo mismo. Veamos por qué.

A tenor de lo establecido cuando se vio el tercer supuesto, la potencia podría incrementarse aumentando el valor de α, incrementando el tamaño de muestra, disminuyendo σ, incrementando la amplitud de la diferencia que queremos detectar $|\theta - \theta_0|$, o sometiendo a comprobación hipótesis unidireccionales. El procedimiento más utilizado es aumentar el tamaño de muestra. Esta estrategia tiene sus riesgos, ya que tamaños de muestra muy elevados harán que detectemos magnitudes de DIF muy pequeñas, es decir, que obtengamos significación estadística en ausencia de significación práctica. En esta situación, disponer de una medida de la magnitud del DIF evitará que descartemos ítems perfectamente válidos. De otro lado, tamaños de muestra muy pequeños hacen disminuir de forma drástica la potencia de prueba (Fidalgo, Mellenbergh y Muñiz, 1999; Muñiz, Hambleton y Xing, 2001; Parshall y Miller, 1995). De nuevo nos podemos proteger de los estragos que el tamaño de muestra provoca en las pruebas de significación estadística, empleando medidas de la magnitud del DIF que proporcionan buenas estimaciones aún con tamaños muestrales pequeños. Así por ejemplo, para el procedimiento MH, Fidalgo, Mellenbergh y Muñiz (1999) encuentran que las estimaciones de la magnitud del DIF obtenidas mediante el cociente de razones común MH ($\hat{\delta}_{MH}$) apenas varían entre tamaños de muestra de 1,000 y 200 personas por grupo, con valores promedio de 1.51 y 1.53, respectivamente. Sin embargo, la potencia del estadístico ji-cuadrado MH para rechazar correctamente la H_0 de ausencia de DIF, sí se vio fuertemente afectada por el tamaño de muestra.

Existe además otra razón de índole estadística para tomar en consideración el tamaño del efecto en la que raramente se repara: que la tasa de error de Tipo I se incrementa conforme lo hace el número de pruebas de significación estadística aplicadas, aunque la tasa de error de Tipo I en cada una de estas pruebas sea constante. Suena raro, pero es así. Supongamos que decidimos incrementar nuestro patrimonio con la ayuda de 10 monedas trucadas. Nosotros siempre apostamos cara, a sabiendas de que existe una probabilidad de 0.95 de obtener tal resultado en cada una de ellas. Por lo tanto, la probabilidad de equivocarnos en cada tirada con cada una de ellas es de 0.05. Sin embargo, como es obvio, la probabilidad de que perdamos alguna vez en cada tirada se incrementará cuantas más apuestas realicemos. Es decir, que es más probable que salga alguna cruz cuantas más monedas lancemos al aire.

De igual forma, la probabilidad de obtener al menos un resultado significativo por azar (error de Tipo I) se incrementa conforme lo hace el número de pruebas aplicadas, aunque en cada una de ellas sea constante y bajo (habitualmente 0.05 o 0.01). Veámoslo más formalmente. Si α es la probabilidad de cometer un error de Tipo I en una prueba, llamémosle α_c de ahora en adelante, $1-\alpha_c$ será la probabilidad de no cometer ningún error de Tipo I. Como sabemos, la probabilidad de ocurrencia conjunta de sucesos independientes es igual al producto de sus probabilidades, luego la probabilidad de no cometer errores de Tipo I en *k* pruebas independientes será igual a $(1-\alpha_c)_1 (1-\alpha_c)_2 \ldots (1-\alpha_c)_k = (1-\alpha_c)^k$, y la probabilidad de cometer *algún* error de Tipo I será $\alpha_F = 1- (1-\alpha_c)^k$. Como se ve, cuando aplicamos *k* pruebas independientes, todas con el mismo riesgo α_c, la probabilidad de cometer un error de Tipo I en esa familia de pruebas es igual a $\alpha_F = 1- (1-\alpha_c)^k$, y no α_c. Esto, dicho en cristiano, significa que si evaluamos uno a uno los 40 ítems de un test para saber si están sesgados contra los negros e hispanos frente a los anglosajones, deberemos realizar 80 pruebas de significación estadística, y es más que probable que alguno de estos ítems sean identificados estadísticamente con DIF, aunque no funcionen diferencialmente.

Una posible solución es ajustar el nivel de significación mediante la desigualdad de Bonferroni, que establece que la probabilidad conjunta de 2 o más sucesos nunca puede exceder la suma de sus probabilidades individuales, es decir, que $\alpha_F \leq k\alpha_c$. Para ajustar la tasa de error, y que se cumpla la desigualdad de Bonferroni, el nivel de significación a utilizar en cada uno de las *k* pruebas, denotémoslo por α_c^*, debe ser igual a: $\alpha_c^* = \alpha_c / k$. Así, si aplicásemos 5 pruebas de significación y quisiésemos mantener en el conjunto de dichas 5 pruebas una tasa de error del 0.05, debiéramos utilizar un nivel de significación del 0.01 (0.05 / 5) en cada una de ellas. De esta forma se cumplirá la desigualdad de Bonferroni,

$$\alpha_F \leq k\ \alpha_c^* = k\ (\alpha_c / k) = \alpha_c$$

Esta solución, que raramente se ha utilizado (Ferreres, González-Romá y Gómez-Benito, 1999), exige utilizar niveles de significación muy bajos en cada una de las pruebas (véase la Tabla 2) lo que hará disminuir la potencia de prueba, y para ciertos tipos de ítems ya se halla bastante comprometida (téngase presente en este contexto lo ilustrado en la Tabla 1). Por ello se sigue la estrategia que venimos defendiendo: utilizar la significación práctica como guía de aceptación de la significación estadística. Los criterios utilizados por el 'Educational Testing Service' (Zieky, 1993, pág. 342) para clasificar a los ítems sometidos a análisis de DIF ejemplifican excelentemente esta estrategia. Criterios similares a la hora de evaluar el DIF con el SIBTEST han sido aconsejados por Roussos y Stout (1996, p. 220).

Tabla 2
Tasa de error por familia de pruebas (α_F), y tasa de error a utilizar en cada prueba de significación tras aplicar la corrección de Bonferroni (α_c^*), siendo *k* el número de pruebas

k	$\alpha_F = 1- (1-\alpha_c)^k$	$\alpha_c^* = \alpha_c / k$
1	0.05	0.05
5	0.23	0.01
10	0.40	0.005

Segunda consideración

Características de los ítems y de las poblaciones

Uno de los objetivos para los que con mayor frecuencia se han empleado los estudios de simulación ha sido el de conocer de forma cierta el efecto que sobre las técnicas de detección del DIF tienen las características de los ítems o la distribución en las poblaciones de la variable medida por el test. Como en dichos estudios la selección de los niveles de las variables manipuladas es en general arbitraria o, por mejor decir, no aleatoria, lo que en la terminología del diseño experimental se denomina un modelo de efectos fijos, la generalización de los resultados debe restringirse exclusivamente a los niveles de las VIs seleccionados. No obstante, la gran abundancia de estudios de esta clase, y el hecho de que presenten resultados bastante coincidentes, nos ofrece una gama suficiente de conocimientos para orientarnos en la mayoría de las situaciones que se nos pueden presentar en la realidad. Sin embargo, en los estudios aplicados se toman en consideración los hallazgos de los estudios de simulación con menor frecuencia de la debida.

Pongamos un ejemplo. Sabemos que la presencia de diferentes distribuciones en la variable medida por el test entre los grupos analizados (lo que se denomina impacto) afecta de forma considerable, y en modo diverso, a las distintas técnicas para evaluar el DIF. Por consiguiente, lo primero que debiéramos hacer antes de proceder a evaluar el DIF es determinar si existe impacto entre las poblaciones de interés. Una simple comparación de medias realizada sobre la variable medida por el test utilizando la *t* de Student como estadístico de contraste bastará. Téngase de nuevo presente que tamaños muestrales grandes llevarán a considerar estadísticamente significativas diferencias mínimas entre los grupos, por lo que será recomendable utilizar algún estimador del tamaño del efecto. Pardo y San Martín (1994) ofrecen buenas indicaciones al respecto. En el caso de que no se cumpliesen los supuestos necesarios para aplicar la prueba *t*, podemos recurrir a la prueba no paramétrica de Mann-Whitney. Retomando la cuestión, conocer si existe impacto nos ayudará a decidir entre técnicas estadísticas alternativas. Así, entre aplicar el procedimiento Mantel-Haenszel o el SIBTEST en presencia de impacto, debiéramos elegir, a priori, el SIBTEST dada su mayor robustez en esas condiciones (Roussos y Stout, 1996b).

De igual forma, las características de los ítems debieran guiarnos a la hora de planificar los análisis estadísticos a realizar. Concretemos. Sabemos que, a igualdad de condiciones, el procedimiento MH detecta peor los ítems muy difíciles o poco discriminativos, es decir, tiene una menor potencia de prueba con ítem de esas características. Por lo tanto, en un intento por aumentar la potencia de prueba, al analizar estos ítems mediante pruebas de significación estadística podemos utilizar niveles de significación mayores (p. ej. 0.05) que para el resto de los ítems (p. ej. 0.01).

Tercera consideración

Concordancia entre procedimientos

Dado que existen multiplicidad de métodos para detectar el DIF (Fidalgo, 1996; Fidalgo y Muñiz, en prensa; Hidalgo y Gómez, 1999; Hidalgo y López Pina, 2000; Penfield y Lam, 2000), cada uno de los cuales presentan peculiaridades que pueden hacerlos más indicados para unos datos o tipos de ítems que otros, el usarlos de forma complementaria puede ser la mejor alternativa. Esto, sin embargo, no es nada sencillo. ¿Qué se entiende por complementario? ¿Qué exigencias y riesgos tiene esta estrategia? Como acostumbramos, antes de nada, un ejemplo. Supongamos que hemos aplicado dos procedimiento para evaluar el DIF sobre los mismos datos: el procedimiento MH y la estandarización. De resultas de ello, hemos identificado 5 ítems con ambos procedimientos, más 2 ítems que se ha detectado sólo con el MH, y 3 más sólo con la estandarización. En ausencia de criterios tenemos dos estrategias para decidir qué ítems eliminar del test. Primera, eliminar todos aquellos ítems que hayan sido detectados por algún método, es decir, eliminaríamos los 10 ítems detectados. Esta estrategia, tiende a maximizar el error de Tipo I, es decir, a aumentar la probabilidad de que eliminemos ítems que no presentan DIF. Además, esta probabilidad aumentará a medida que aumente el número de pruebas aplicadas, ya que se irán añadiendo los falsos positivos obtenidos con cada nuevo método aplicado. Segunda estrategia, dar por identificaciones correctas sólo a aquellos ítems que hayan sido identificados por todos los métodos, es decir, en nuestro ejemplo eliminaríamos sólo 5 ítems. Esta postura tiende a maximizar el error de Tipo II, es decir, que aumenta la probabilidad de pasar por alto ítems que presentan DIF. Como en el caso anterior, dicha probabilidad aumentará a medida que aumente el número de pruebas aplicadas.

En consecuencia, en contra de lo que pudiera parecer al pronto, el objetivo de aplicar varios métodos no debe ser obtener evidencia de DIF en función del número de métodos diferentes por los que ha sido identificado el mismo ítem, sino encontrar patrones de resultados indicativos de DIF que puedan ser interpretados coherentemente a partir del conocimiento teórico que de dichas técnicas se tiene. A no ser, claro, que como hemos indicado anteriormente, queramos maximizar el error de Tipo I o el error de Tipo II. Por seguir con el ejemplo puesto, a la hora de evaluar los resultados obtenidos deberíamos tener en consideración cosas como que, bajo ciertas condiciones, el error típico de la diferencia de proporciones estandarizadas (DPE) es mayor cuanto menor sea la varianza de la habilidad medida por el test, y que, por el contrario, el error típico del estadístico $\hat{\alpha}_{MH}$ se minimiza en esas mismas condiciones, como resultado de las diferentes métricas en que se expresan ambos índices (Zwick, 1997); o el mayor efecto que la dificultad del ítem tiene sobre el estadístico DPE (Dorans y Holland, 1993). Queda al criterio del analista sopesar las ventajas e inconvenientes de aplicar varios métodos de evaluación del DIF, dadas las características de sus datos, de las propiedades estadísticas de las técnicas a aplicar, y del conocimiento que tenga de ellas.

Corolario

La toma de decisiones a la hora de someter a comprobación estadística hipótesis sobre el DIF siempre debe hacerse poniendo en conexión las implicaciones que de tales decisiones se derivan con los costes que uno está dispuesto a asumir.

En el intento de demostrar la pertinencia de la conclusión obtenida, a lo largo del artículo se ha ilustrado el coste que, en términos de error de Tipo I y de Tipo II, tienen varias de las decisiones con más frecuencia aplicadas: la utilización de pruebas de evaluación del DIF unidimensionales en contextos multidimensionales, la evaluación del DIF mediante un primer y único análisis, la ausencia de medidas del tamaño del efecto, la elección de los niveles de significación convencionales, y la utilización de varios procedimientos para evaluar el DIF, entre otras. Lo dicho han sido, por lo tanto, consideraciones en torno a la evaluación estadística del DIF, de forma que incluso cuando se ha tratado sobre la significación práctica se ha hecho desde la perspectiva estadística del tamaño del efecto. Hemos decidido restringir la discusión a estos temas, a sabiendas de que un estudio de DIF es mucho más que la aplicación exploratoria de unos análisis estadísticos. La corrección de los análisis estadísticos y su correcta interpretación es, sin embargo, esencial. Además, sabemos que alguno de los puntos tratados cobran su máximo sentido dentro del contexto de una evaluación del DIF ítem a ítem, y que otras alternativas son posibles (véase Fidalgo y Muñiz, en prensa; Gierl, Bisanz, Bisanz, Boughton, y Khaliq, 2001). Finalmente, el lector interesado puede encontrar un ejemplo ilustrado empíricamente de parte de lo dicho aquí en Fidalgo, Ferreres y González-Romá (2001).

Si empezamos con Heráclito, buena cosa es acabar con una cita suya que viene como anillo al dedo a los estudios empíricos sobre DIF: φύσις κρύπτεσθαι φιλεῖ (la auténtica naturaleza de las cosas suele estar oculta).

Referencias

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29* (1), 67-91.

Camilli, G. y Shepard, L.A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Cuesta, M. (1996). Unidimensionalidad. En J. Muñiz (Coord.), *Psicometría* (pp. 239-292). Madrid: Universitas.

Dorans, N.J. y Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. En W.P. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: LEA.

Ferreres, D., González-Romá, V. y Gómez-Benito, J. (2000). Comparación del estadístico Mantel-Haenszel y la regresión logística en el funcionamiento diferencial de los ítems en dos pruebas de aptitud intelectual en un contexto bilingüe. *Psicothema, 12* (2), 214-219.

Fidalgo, A.M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), *Psicometría* (pp. 370-455). Madrid: Universitas.

Fidalgo, A.M., Ferreres, D. y González-Romá, V. (2001, Septiembre). *Evaluación del funcionamiento diferencial de los ítems en datos empíricos*. Comunicación presentada en el VII Congreso de Metodología de las Ciencias Sociales y de la Salud, Madrid.

Fidalgo, A.M., Mellenbergh, G.J. y Muñiz, J. (1998). Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems. *Psicothema, 10*, 219-228.

Fidalgo, A.M., Mellenbergh, G.J. y Muñiz, J. (1999). Efectos de la aplicación del procedimiento Mantel-Haenszel en una etapa, en dos etapas e iterativamente sobre los estadísticos Mantel-Haenszel. *Psicológica, 20*, 227-242.

Fidalgo, A.M., Mellenbergh, G.J. y Muñiz, J. (2000). Effects of Amount of DIF, Test Length, and Purification Type on Robustness and Power of Mantel-Haenszel Procedures. *Methods of Psychological Research Online, 5* (3), 43-53. Available: http://www.mpr-online.de

Fidalgo, A.M., y Muñiz, J. (en prensa). *Líneas de investigación actuales sobre el funcionamiento diferencial de los ítems*. Metodología de las Ciencias del Comportamiento.

Flowers, C.P., Oshima, T.C., y Raju, N.S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23* (4), 309-326.

Gierl, M.J., Bisanz, J., Bisanz, G.L., Boughton, K.A. y Khaliq, S.N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement test. *Educational Measurement: Issues and Practice*, *20*, 26-36.

Hattie, J. (1985). Metodology Review: assessing unidimensionality of test and items. *Applied Psychological Measurement, 9* (2), 139-164.

Hidalgo, M.D. y Gómez, J. (1999). Técnicas de funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento, 2*, 167-182.

Hidalgo, M.D. y Lopez-Pina, (2000). Funcionamiento diferencial de los ítems: presente y perspectivas de futuro. *Metodología de las Ciencias del Comportamiento, 2*, 167-182.

Kok, F.G. (1988). Item bias and test multidimensionality. En R. Langeheine y J. Rost (Ed.), *Latent trait and latent class models* (pp. 263-274). New York: Plenum.

Mazor, K.M., Hambleton, R.K. y Clauser, B.E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22* (4), 357-367.

Mazor, K.M., Kanjee, A. y Clauser, B.E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131-144.

Miller, M.D. y Linn, R.L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*, 205-219.

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.

O'Neill, K.A. y McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. En W.P. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: LEA.

Oshima, T.C., Raju, N.S. y Flowers, C.P., (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*, 253-272.

Parshall, C.G., y Miller, T.R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32* (3), 302-316.

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.

Rosnow, R.L. y Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Roussos, L. y Stout, W. (1996a). A multidimensionality-based DIF paradigm. *Applied Psychological Measurement, 20* (4), 355-371.

Roussos, L. y Stout, W. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Ryan, K.E. y Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14* (1): 73-90.

Shealy, R. y Stout, W. (1993). A model-based standarization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Stout, W., Li, H., Nandakumar, R. y Bolt, D. (1997). MULTISIB: a procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21* (3), 195-213.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. En W.P. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: LEA.

Zwick, R. (1997). The effect of adaptative administration on the variability of the Mantel-Haenszel measure of differential item functioning. *Educational and Psychological Measurement, 57*, 412-121.