FUNCIONAMIENTO DIFERENCIAL DEL ITEM DIF

Compilación CEO

Abstract. Tests or standardized measuring instruments has become an indispensable tool in social sciences, such as Psychology, Sociology and the Education. For that reason he is fundamental that the professionals use who them assure that these instruments work in the same way in the different existing groups based on sociodemographic variables like sex, the culture, the ethnic group, etc. When some of the items diferencialmente works for the different groups operation speech differential of the item (DIF). Indeed the objective of the present work concentrates in presenting/displaying a brief introduction to the concept of DIF and the existing methods for its detection.

Los tests o instrumentos de medida estandarizados se han convertido en una herramienta indispensable en las ciencias sociales, tales como la Psicología, la Sociología y la Educación. Por ello es primordial que los profesionales que los emplean se aseguren de que estos instrumentos funcionan del mismo modo en los distintos grupos existentes en función de variables socio-demográficas como el sexo, la cultura, la etnia, etc. Cuando alguno de los ítems funciona diferencialmente para los distintos grupos se habla de funcionamiento diferencial del ítem (DIF)[1]. Precisamente el objetivo del presente trabajo se centra en presentar una breve introducción al concepto de DIF y a los métodos existentes para su detección.

[1] El funcionamiento diferencial se evalúa con el estadístico Mantel-Haenszel

Introducción

Desde la aparición de los primeros tests en Estados Unidos para la selección de personal en empresas, la admisión de estudiantes en la universidad, de soldados en el ejército, y en otros muchos sectores públicos y privados, los tests o instrumentos de medida estandarizados se han convertido en una herramienta indispensable en las ciencias sociales, tales como la Psicología, la Sociología y la Educación. Concretamente, en estos ámbitos, se utilizan una gran variedad de instrumentos de medida para obtener información acerca de habilidades, opiniones, actitudes, etc. de las personas y grupos sociales. En este sentido es primordial que los profesionales que emplean tests se cercioren de que éstos funcionan del mismo modo en los distintos grupos existentes en función de variables socio-demográficas, como el sexo, la cultura, la etnia, etc., ya que cualquier instrumento de medida tiene que ser objetivo en su medición, es decir, ha de garantizar resultados idénticos en sujetos que tienen el mismo nivel en el atributo medido sea cual sea el grupo de pertenencia. En el caso de que alguno de los ítems que componen el test funcione distintamente a favor (o en contra) de sujetos que tienen habilidad similar, se habla de funcionamiento diferencial del ítem (DIF, Differential Item Functioning), término introducido en 1988 por Holland y Thayer.

Como consecuencia del importante papel que juegan los tests en la toma de decisiones tanto a nivel social y jurídico como psicológico y educativo, los estudios de DIF constituyen uno de los campos de investigación que más interés ha suscitado en el ámbito psicométrico durante las últimas décadas; evidencia de ello es la existencia de gran cantidad de publicaciones que analizan la imparcialidad de los tests con respecto a distintos grupos de sujetos. Por ello, el objetivo de este trabajo se centra en presentar, a nivel teórico**, una breve introducción al concepto de DIF** y a los métodos existentes para su detección, aspectos claves a tener en cuenta en la práctica sociológica, psicológica y educativa.

El funcionamiento diferencial del ítem Definición del DIF Un determinado ítem o test presenta DIF si se comporta diferencialmente para individuos o grupos comparables, que difieren en lengua nativa, género, etnia, cultura, o cualquier otra variable que pueda constituir una fuente sistemática de variación ajena al rasgo medido por la prueba en cuestión, entendiendo por comparables aquellos grupos de sujetos que poseen el mismo nivel en la característica o rasgo medido por el test (Gómez e Hidalgo, 1997). En otras palabras, un ítem presenta DIF **cuando dos grupos comparables presentan una probabilidad distinta de responder con éxito** o en determinada dirección dicho ítem en función del grupo al que pertenezcan; de este modo uno de los grupos presentará una ventaja relativa respecto al otro.

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 sujetos estándar respecto al cual se compara el grupo focal. Aunque en la mayoría de los casos se trabaja con un grupo focal y uno de referencia, cabe la posibilidad de contemplar más de un grupo focal, que se comparan con el mismo grupo de referencia. DIF, impacto y sesgo.

Un aspecto a tener en cuenta en el estudio de la equidad métrica de los instrumentos de medida se refiere a la distinción entre DIF, impacto y sesgo. Hasta el momento ya hemos señalado que un ítem presenta DIF cuando, a iguales niveles del rasgo o habilidad medida, desfavorece a un grupo de sujetos frente a otro. Sin embargo, hablaremos de impacto en el caso de que las diferencias encontradas entre grupos sean causa de una diferencia real en la variable medida (Ackerman, 1992); en términos de probabilidades, un ítem presenta impacto si la probabilidad de responderlo correctamente es mayor para aquel grupo que

realmente es superior en la habilidad medida, mientras que la probabilidad de acertarlo es la misma para todos aquellos individuos con un mismo nivel en el rasgo, independientemente del grupo al cual pertenecen.

Finalmente, a diferencia de los estudios de detección de DIF, los estudios de sesgo dan un paso más adelante; pretenden encontrar una explicación lógica a las causas que subyacen en el modo de funcionar diferencialmente de algunos ítems entre grupos. En el caso de los estudios de DIF se emplean técnicas estadísticas para su detección, pero en los estudios de sesgo son los sociólogos, psicólogos, antropólogos y educadores profesionales los que analizan minuciosamente, en relación al constructo medido por el test, cada uno de los ítems detectados con DIF para poder concluir si presentan o no sesgo.

Tipos de DIF

Mellenbergh (1982) 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 sujetos. En el denominado uniforme no existe interacción entre el nivel en el rasgo medido y la pertenencia a un determinado grupo, es decir que la probabilidad de responder correctamente al ítem en cuestión 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 dicha interacción, por lo que la probabilidad de cada grupo de responder correctamente al ítem no es la misma a lo largo de los distintos niveles del rasgo medido. Desde la teoría de respuesta al ítem se propone el concepto de curva característica del ítem (CCI), de gran utilidad para entender gráficamente los distintos tipos de DIF.

La CCI relaciona la probabilidad de acertar el ítem con el nivel de los sujetos en la variable medida. De este modo, un ítem no presenta DIF si su curva característica para el grupo focal y para el grupo de referencia coinciden, muestra DIF uniforme si las respectivas CCIs no se cruzan en ninguno de los niveles en la variable

medida, y presenta DIF no uniforme si en algún punto éstas sí que se cruzan. A continuación se presentan tres ejemplos gráficos que esperamos faciliten la comprensión del concepto y tipología del DIF:

Métodos de detección del DIF

Dada la importancia de la detección del DIF para asegurar la equidad métrica de los instrumentos de medida, existen en la actualidad una extensa variedad de técnicas de análisis para detectar el DIF. En este apartado presentamos una aproximación a dichos procedimientos sin entrar en la explicación detallada de cada uno de ellos, ya que no es el objetivo central del presente artículo.

Aquel lector interesado en profundizar en los procedimientos aquí nombrados puede consultar las revisiones de Gómez e Hidalgo (1997), Hidalgo y Gómez (1999) e Hidalgo y López-Pina (2000) en castellano, y Millsap y Everson (1993) y Potenza y Dorans (1995) en inglés.

Métodos incondicionales vs condicionales

Los métodos incondicionales de detección del DIF se basan en las diferencias en la dificultad del ítem y se caracterizan por igualar a los sujetos respecto al nivel en la habilidad o rasgo medido, de ahí que los grupos no pueden considerarse comparables y, por tanto, son procedimientos ya descartados. Las técnicas más destacadas son el análisis de la varianza (ANOVA), el análisis delta-plot y la correlación biserial-puntual, métodos actualmente poco utilizados para la detección del DIF ya que presentan una tasa de detecciones correctas muy baja y ocasionan una elevada tasa de falsos positivos (ítems que, sin presentar DIF, son detectados con DIF).

Los métodos condicionales, a diferencia de los incondicionales, permiten trabajar con grupos comparables porque igualan los grupos respecto al rasgo medido. Métodos de invarianza condicional observada vs no observada Las técnicas condicionales pueden igualar los grupos tomando en consideración diferentes aspectos, que definen los dos tipos de procedimientos: métodos de invarianza condicional observada y de invarianza condicional no observada (Millsap y Everson, 1993).

Los primeros utilizan como criterio de comparabilidad entre los grupos una variable observada, normalmente la puntuación total de los sujetos en el test, sin especificar ningún modelo de medida; cabe citar la prueba (2, los modelos loglineales y logit, el estadístico Mantel-Haenszel y la regresión logística, entre otros. Sin embargo, en el segundo conjunto de procedimientos la variable de igualación es una variable latente; se formula un modelo basado en la teoría de respuesta al ítem o en el análisis factorial confirmatorio, y se comprueba si los parámetros estimados en dicho modelo se mantienen invariantes para los distintos grupos.

Métodos para ítems dicotómicos vs politómicos

Las pruebas citadas hasta el momento son aplicables a ítems de respuesta dicotómica, es decir, aquellos ítems que solamente permiten dos categorías de respuesta (p. ej: Sí/No y Acierto/Error), pero en muchas ocasiones, y actualmente ésta es la tendencia, los tests presentan un formato de respuesta con más de dos categorías (p. ej: escalas tipo Likert) denominados ítems politómicos. En este último caso, han surgido diversas técnicas de detección del DIF, generalmente adaptando los procedimientos utilizados en ítems dicotómicos y proponiendo extensiones para ítems de respuesta politómica. Entre ellos, podemos destacar los métodos basados en la teoría de respuesta al ítem, las generalizaciones del estadístico Mantel-Haenszel, las extensiones de la regresión logística, etc.

Métodos de purificación

Un problema de igualar los grupos utilizando como criterio de comparación la variable que mide el test, ya sea observada o latente, hace referencia a que dicho criterio de igualación está contaminado por la presencia de los ítems que muestran DIF y que forman parte del criterio junto a los ítems sin DIF. En este sentido, para reducir el efecto producido por los ítems con DIF, se han propuesto algunas técnicas de purificación que iterativamente eliminan del criterio aquellos ítems que en etapas previas presentan DIF. Por ejemplo, Holland y Thayer (1988) emplean el método de purificación bietápica para el estadístico Mantel-Haenszel, Gómez y Navas (1996) aplican la purificación paso a paso a la regresión logística dicotómica, y Hidalgo y Gómez (2003) a la regresión logística politómica, entre otros.

Discusión

A lo largo de estas líneas hemos pretendido concienciar a los profesionales de las ciencias humanas y sociales de la importancia de la detección de ítems que funcionan diferencialmente para diversos grupos, y así asegurarse de que los instrumentos de medida utilizados no están aventajando o desfavoreciendo a

alguno de los grupos sometidos a estudio. Como amenaza que supone el DIF para la validez de los instrumentos de medida, los estudios de su detección deberían suponer una fase añadida tanto al proceso de evaluación de los instrumentos de medida ya estandarizados como al desarrollo de nuevos tests, recomendación señalada en los últimos Standards for Educational and Psychological Testing (APA, AERA, NCME, 1999).

En línea con esta concienciación progresiva, hemos presentado en este trabajo una introducción al concepto de DIF y sus distintos tipos, y una breve caracterización de las principales aportaciones metodológicas para la detección del DIF. Esperamos que ello favorezca y facilite el interés por una mayor profundización en el tema y una utilización óptima de los instrumentos de captura de datos en el campo de las ciencias sociales.

Referencias

1.-ACKERMAN, Terry. “A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective”. En: Journal of Educational Measurement, 29, 1. Washington: 1992, p. 67-91.

2.-AMERICAN PSYCHOLOGICAL ASSOCIATION, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION y NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. Standards for educational and psychological testing. Washington: American Psychological Association, 1999.

3.-GÓMEZ BENITO, Juana y HIDALGO MONTESINOS, María Dolores. “Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica”. En: Anuario de Psicología, 74, 3. Barcelona: 1997, p. 3-32.

4.-GÓMEZ BENITO, Juana y NAVAS ARA, María José. “Detección del funcionamiento diferencial del ítem: Purificación paso a paso de la habilidad”. En: Psicológica, 17. Valencia: 1996, p. 397-411.

5.-HIDALGO MONTESINOS, María Dolores y GÓMEZ BENITO, Juana. “Técnicas de detección del funcionamiento diferencial en ítems politómicos”. En: Metodología de las Ciencias del Comportamiento, 1, 1. Murcia: 1999, p. 39-60.

6.- HIDALGO MONTESINOS, María Dolores y GÓMEZ BENITO, Juana. “Test purification and the evaluation of differential item functioning with multinomial logistic regression”. En: European Journal of Psychological Assessment, 19, 1. Göttingen: 2003, p. 1-11.

7.-HIDALGO MONTESINOS, María Dolores y LÓPEZ-PINA, José Antonio. “Funcionamiento diferencial de los ítems: Presente y prespectivas de futuro”. En: Metodología de las Ciencias del Comportamiento, 2, 2. Murcia: 2001, p. 167-182.

8.-HOLLAND, Paul y THAYER, Dorothy. “Differential item performance and the Mantel-Haenszel procedure”. Test Validity. Hillsdale: LEA, 1988, p. 129-145.

9.-MELLENBERGH, Gideon. “Contingency table models for assesing item bias”. En: Journal of Educational Statistics, 7. Washington: 1982, p. 105-118.

10.-MILLSAP, Roger y EVERSON, Howard. “Methodology review: Statistical approaches for assessing measurement bias”. En: Applied Psychological Measurement, 17. Thousand Oaks: 1993, p. 297-334.

11.-POTENZA, Maria y DORANS, Neil. “DIF assessment for politomously scored items: A framework for classification and evaluation”. En: Applied Psychological Measurement, 19. Thousand Oaks: 1995, p. 23-37.)