I Congreso Internacional de Investigación y Práctica Profesional en Psicología XVI Jornadas de Investigación Quinto Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires, 2009.

Error de tipo I de las Pruebas de Breslow-Day y reglas combinadas aplicadas a la detección del funcionamiento diferencial del ítem. Un estudio sobre un test corto en presencia de impacto.

Aguerri, María Ester, Picón Janeiro, Jimena, Blum, G. Diego, Abal, Facundo Juan Pablo, Lozzia, Gabriela y Galibert, María Silvia.

ERROR DE TIPO I DE LAS PRUEBAS DE BRESLOW-DAY Y REGLAS COMBINADAS APLICADAS A LA DETECCIÓN DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM. UN ESTUDIO SOBRE UN TEST CORTO EN PRESENCIA DE IMPACTO

Aguerri, María Ester; Picón Janeiro, Jimena; Blum, G. Diego; Abal, Facundo J.P.; Lozzia, Gabriela; Galibert, María Silvia
Agencia Nacional de Promoción Científica y Tecnológica - Facultad de Psicología, Universidad de Buenos Aires

RESUMEN
En el presente trabajo se analiza la tasa de error de Tipo I de métodos de detección del funcionamiento diferencial del ítem (Differential Item Functioning, DIF) sobre tests de 20 ítems cuando los grupos difieren en la habilidad medida. Las pruebas bajo estudio son las de Breslow-Day, global (BD) y de la tendencia (BDT), y reglas que las combinan con el procedimiento de Mantel-Haenszel (MH). Randall Penfield propuso la Regla de Decisión Combinada (RDC) que considera la aplicación de MH y BDT con el ajuste de Bonferroni. Las reglas combinadas son aptas para detectar todo tipo de DIF pues BD y BDT son aptas para detectar la presencia de DIF no Uniforme y MH es considerado el más potente para la detección del DIF Uniforme. Se consideraron las respuestas simuladas sin DIF a tests de 20 ítems de 100 pares de grupos, de 1,000 sujetos cada uno, que difieren en la habilidad. Analizada la proporción de detección errónea del DIF se obtuvo que de las pruebas de Breslow-Day, sólo la global presentó, al 1% y al 5%, valores próximos a los esperados. MH exhibió tasas de falsos positivos particularmente infladas, situación que explica las altas tasas de las reglas combinadas.

Palabras clave
Funcionamiento diferencial de ítem Pruebas de Breslow-Day Procedimiento de Mantel-Haenszel

ABSTRACT
TYPE I ERROR OF THE BRESLOW-DAY TESTS AND COMBINED RULES APPLIED TO THE DETECTION OF THE DIFFERENTIAL ITEM FUNCTIONING. A STUDY CONDUCTED ON SHORT TESTS IN THE PRESENCE OF IMPACT
The purpose of this study is to analyze the type I error rate of the Differential Item Functioning detection methods in 20-item tests when the groups differ in the ability measured. The tests included in the study are the Breslow-Day Tests, both global (BD) and of trend (BDT), and the rules that combine them with the Mantel-Haenszel procedure (MH). Randall Penfield proposed the Combined Decision Rule (CDR) that considers the application of MH and BDT with Bonferroni's adjustment. The combined rules are suitable for the detection of all types of DIF given that BD and BDT are suitable for the detection of nonuniform DIF and MH is considered the most powerful for the detection of uniform DIF. This study considered the simulated answers without DIF in 20-item tests of 100 pairs of groups, composed of 1,000 individuals each, which differ in ability. After analyzing the DIF erroneous detection rate, it was concluded that of the Breslow-Day tests, only the global neared the expected rates: at 1% and at 5%. MH showed particularly inflated false positives rates, which explains the high rates yielded by the combined rules.

Key words
Differential item functioning Breslow-Day tests Mantel-Haenszel procedure

El análisis del funcionamiento diferencial del ítem (*Differential Item Functioning*, DIF) está incorporado a los estudios actualizados sobre la validez de los instrumentos de medición. DIF es el término estadístico utilizado para describir situaciones en las que personas de un grupo responden al ítem de manera correcta más frecuentemente que personas igualmente capaces de otro grupo (Zumbo, 2007). La existencia de DIF indica la posible presencia de una variable extraña a la habilidad de interés que explica la respuesta al ítem y pone en evidencia la falta de validez del instrumento. Camilli y Shepard (1994) mencionan, entre otros métodos de detección del DIF, los llamados de Tablas de Contingencia. Para estos métodos se consideran igualmente capaces los sujetos con un mismo puntaje total en el test. Básicamente se ocupan de analizar las tablas 2x2 que resultan al considerar el grupo de pertenencia (de Referencia y Focal según la literatura específica) y las posibles respuestas al ítem (correcta o incorrecta) para todos los niveles del puntaje total. Si los sujetos de ambos grupos presentan la misma posibilidad (Odds) de respuesta correcta a lo largo de los niveles del puntaje total se dice que el ítem no presenta DIF, es decir, un ítem no presenta DIF cuando el cociente de las posibilidades (Odds Ratio, OR) es 1 para todos los niveles del puntaje total. Si el OR es constante, y diferente de 1, se dice que el ítem presenta DIF Uniforme. En cambio, si no es constante se dice que el ítem presenta DIF no Uniforme. La prueba de Mantel-Haenszel (MH) y la prueba global de Breslow-Day para la heterogeneidad de los OR (BD) son mencionadas por Camilli y Shepard (1994) para decidir, respectivamente, acerca de la presencia de DIF y de DIF no Uniforme.

Breslow y Day (1980) presentaron pruebas estadísticas que permiten concluir acerca de la heterogeneidad de los OR en el análisis de tablas de 2x2 a lo largo de los estratos de una tercera variable. Estas pruebas fueron propuestas en el marco de estudios sobre cáncer y son aplicables a la detección del DIF no Uniforme. En Aguerri, Galibert, Lozzia y Attorresi (2004) y Aguerri, Galibert, Attorresi y Prieto-Marañón (2009) se aplicó BD al estudio del DIF sobre un test de 20 ítems. Penfield (2003) utilizó la prueba de Breslow-Day de la tendencia en la heterogeneidad de los OR (BDT) sobre un test de 40 ítems.

El procedimiento MH, originalmente presentado por Mantel y Haenszel (1959) para estudios vinculados con el cáncer, permite decidir si el OR es 1 a lo largo de todos los niveles del puntaje total, o no. Holland y Thayer (1988) recomendaron este procedimiento para detectar la presencia o ausencia de DIF. Es una de las pruebas más difundidas para el análisis del DIF pues es de fácil comprensión y puede aplicarse mediante paquetes estadísticos como el SPSS. Por otra parte, Mazor, Clauser y Hambleton (1994) propusieron el procedimiento de Mantel-Haenszel modificado (MHmo) que consiste en realizar tres análisis del DIF mediante el procedimiento MH con un mismo nivel de significación α. Mazor et al. (1994) mostraron sobre un test de 75 ítems que el procedimiento MHmo es más potente que MH frente al DIF no Uniforme.

Dado que MH es el más potente para la detección del DIF Uniforme y las pruebas de Breslow-Day son aptas para identificar al DIF no Uniforme, resulta de interés aplicar reglas que los combinen pues se espera detecten ambos tipos de DIF. Cada una de estas reglas se basa en la decisión de dos pruebas de hipótesis realizadas con nivel de significación α. Éstas son: MHoBD, que señala con DIF a los ítems así identificados por MH o BD, y MHoBDT, basada en la detección de MH o BDT. Por otra parte, Penfield (2003) propuso la Regla de Decisión Combinada (RDC) según la cual un ítem presenta DIF si MH o BDT lo detectan, cada una de estas pruebas con nivel $\alpha/2$. Tal modificación del nivel de significación se denomina *ajuste de Bonferroni.* Penfield (2003) concluyó que RDC tiene resultados superiores a los de la regresión logística y del crossing SIBTEST en cuanto al error de Tipo I.

Acerca del procedimiento MH es mucho lo que se ha estudiado, pero no ocurre lo mismo con las pruebas de Breslow-Day y las reglas combinadas. No se conocen estudios acerca de la tasa de detección errónea del DIF de la prueba de Breslow-Day de la tendencia ni de las reglas combinadas sobre tests cortos cuando los grupos difieren en la habilidad medida (presencia de impacto). En tal situación Aguerri, Picón-Janeiro, Abal y Galibert (2008) estudiaron la tasa de error de Tipo I sobre tests largos y de longitud

moderada. Cuando los grupos no difieren en la habilidad medida Aguerri et. al (2008, Julio) analizaron la tasa de detección errónea para tests cortos, de longitud moderada y largos. El objetivo del presente trabajo es estudiar la tasa de error de Tipo I de las pruebas de Breslow-Day y reglas combinadas aplicadas a la detección del DIF sobre tests cortos en presencia de impacto.

MÉTODO

Se consideraron las respuestas a un test de 20 ítems sobre un diseño similar al de Aguerri et al. (2008, Julio) pero con presencia de impacto. Las respuestas fueron simuladas mediante el programa PARDSIM ® (Yoes, 1997) según el modelo logístico de tres parámetros. El parámetro de aciertos por azar se fijó en 0.20 en todos los casos. Los parámetros de dificultad y discriminación resultaron de combinar cinco niveles del parámetro de dificultad (-1.5, -1, 0, 1 y 1.5) con cuatro niveles del parámetro de discriminación (0.25, 0.60, 0.90 y 1.25). En todos los casos los grupos bajo estudio fueron de tamaño 1,000. La situación de impacto se introdujo considerando que los sujetos del grupo Focal pertenecen a una población normal con habilidad media -1 y desvío 1, mientras que los del grupo de Referencia pertenecen a una población cuya habilidad se distribuye como una normal estándar. Para cada situación se simularon 100 pares de patrones de respuesta con los mismos parámetros generadores en ambos grupos, esto es, sin DIF. Mediante el Programa Computacional Bday (Prieto-Marañón, 2005) se estudió el DIF en cada par de grupos con BD, BDT, MH, MHoBD, MHoBDT, MHmo y RDC y se registró si detectaron, o no, DIF al 1% y al 5%. El programa Bday aplica el procedimiento MH, como el PROC FREQ de SAS (Statistical Analysis System, 1989), en una sola etapa y sin la corrección por continuidad. Finalmente se registró la proporción de DIF erróneamente detectado, tasa de falsos positivos sobre 100 repeticiones, para cada uno de los métodos bajo estudio.

RESULTADOS

La proporción de falsos positivos para los tests de 20 ítems, cuando se trabajó al 1% fue 0.0130 para BD, 0.0270 para BDT, 0.1215 para MH, 0.1335 para MHoBD, 0.1460 para MHoBDT, 0.1585 para MHmo y 0.1070 para RDC. Las proporciones correspondientes cuando se trabajó al 5% fueron: 0.0700, 0.0780, 0.2540, 0.3100, 0.3180, 0.3460 y 0.2255. Sólo la prueba global de Breslow-Day satisfizo la condición liberal establecida por Bradley (1978) al 1% y al 5%, pues la proporción de detección errónea resultó comprendida entre $0.5*\alpha$ y $1.5*\alpha$. No ocurrió lo mismo con la prueba de Breslow-Day de la tendencia que presentó una tasa de falsos positivos particularmente inflada al 1%. El procedimiento MH presentó una tasa de detección errónea muy alta, tanto al 1% como al 5%. Las reglas combinadas, RDC incluida, y MHmo exhibieron tasas de error de Tipo I altamente infladas, esto se explica por la marcada propensión de MH a detectar erróneamente DIF en estas condiciones de diseño. Entre ellas, MHmo presentó los valores más altos y RDC los más bajos.

DISCUSIÓN

La tasa de error de Tipo I resultó inflada para la prueba de Breslow-Day de la tendencia en la heterogeneidad de los Odds Ratio, para el procedimiento estándar de Mantel-Haenszel y, en consecuencia, para las reglas que involucran a una de tales pruebas o a las dos. Sólo la prueba global de Breslow-Day presentó tasas próximas a las esperadas. Esta situación favoreció el desempeño, en cuanto a la tasa de detección errónea, de la regla que combina MH con BD frente a la que combina MH con BDT. Entre las reglas combinadas, la que presentó la tasa de detección errónea menos inflada fue, como era de esperar, la Regla de Decisión Combinada de Penfield pues considera el *ajuste de Bonferroni* en el nivel de significación. El procedimiento MHmo presentó los valores más altos de detección errónea por basarse en la realización de tres pruebas de hipótesis en lugar de dos como lo requieren las otras reglas. Queda en evidencia que la reducción de la longitud del test afectó a la tasa de detección errónea de la prueba de Breslow-Day de la tendencia pues para los tests de longitud moderada en presencia de impacto presentó valores aceptables (Aguerri et al., 2008). La presencia de impacto afectó el comportamiento de BDT y MH sobre tests cortos, en tanto que cuando no lo hay, presentan tasas de falsos positivos similares a las esperadas (Aguerri et al., 2008, Julio). En base a los resultados obtenidos en este trabajo, y en consonancia con los de Aguerri et al. (2009), puede afirmarse que en tests cortos la prueba global de Breslow-Day es la más recomendable en cuanto al riesgo de cometer error de Tipo I. A partir de estudios de la potencia de las pruebas de Breslow-Day y las reglas combinadas sobre similares características de diseño, se podrá decidir cuál es el método más aconsejable para estudiar el DIF en cuanto al riesgo de no detectar a un ítem que sí tenga DIF.

BIBLIOGRAFÍA

AGUERRI, M.E.; GALIBERT, M.S.; ATTORRESI, H.F. & PRIETO-MARAÑÓN, P. (2009). Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test. Quality and Quantity. International Journal of Methodology, 43, 1, pp. 35-44. Springer, Netherland

AGUERRI, M.E.; GALIBERT, M.S.; LOZZIA, G.S. & ATTORRESI, H.F. (2004). Un estudio acerca del funcionamiento diferencial no uniforme del ítem. Metodología de las Ciencias del Comportamiento. Asociación Española de Metodología de las Ciencias del Comportamiento. Murcia, España. Volumen Especial, 7-10.

AGUERRI, M.E.; PICÓN-JANEIRO, J.; ABAL, F.J.P. & GALIBERT, M.S. (2008). Efecto de la longitud del test en la detección errónea del funcionamiento diferencial del ítem de las pruebas de Breslow-Day y reglas combinadas. Un estudio en presencia de impacto. Trabajo Libre. XV Jornadas de Investigación y Cuarto Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología. UBA. Argentina. 6-9/08/2008. ISSN Nº 1667-6750. Tomo II, pp. 455-456.

AGUERRI, M.E.; PICÓN-JANEIRO, J.; LOZZIA, G.; ABAL, F.J.P.; GALIBERT, M.S. & ATTORRESI, H.F. (2008, Julio) Influence of the test length on the erroneous DIF detection of the Breslow-Day tests and combined rules. Trabajo Libre. III European Congress of Methodology. European Association of Methodology. Oviedo, España. Libro de Resúmenes, p. 64.

BRADLEY, J.V. (1978). Robustness? The British Journal of Mathematical & Statistical Psychology, 31, 144-152.

BRESLOW, N.E. & DAY, N.E. (1980). Statistical Methods in Cancer Research. Volume I. The Analysis of Case-Control Studies. Lyon, France. International Agency for Research on Cancer (IARC Scientific Publication No. 32)

CAMILLI, G. & SHEPARD, L. (1994). Methods for Identifying Biased Test Items. Thousand Oaks: SAGE.

HOLLAND, P.W. & THAYER, D. . (1988). Differential item functioning and the Mantel- Haenszel procedure. En H. Wainer & H.I.Braun (Eds.), Test Validity (pp. 129 -145). Hillsdale, NJ: Lawrence Erlbaum.

MANTEL, N. & HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

MAZOR, K.M.; CLAUSER, B.E. & HAMBLETON, R.K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. Educational and Psychological Measurement, 54, 284-291

PENFIELD, R. (2003). Applying the Breslow-Day test of trend in Odds Ratio heterogeneity to the analysis of nonuniform DIF. The Alberta Journal of Educational Research, Vol. XLIX, 231-243.

PRIETO-MARAÑÓN, P. (2005). Bday: Programa computacional para el estudio del DIF mediante las pruebas de Breslow-Day, los procedimientos de Mantel-Haenszel y reglas combinadas. Inédito.

SAS Institute Inc., (1989). SAS/STAT® User's Guide. Version 6, Fourth Edition, Volume 1, Cary, N.C.: SAS Institute Inc., 943 pp.

YOES, M. (1997). PARDSIM Parameter and Response Data Simulation [Software]. St. Paul, MN: Assessment System Corporation.

ZUMBO, B.D. (2007). Three generations of DIF analysis: considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, 4, 223-233.