XV Jornadas de Investigación y Cuarto Encuentro de Investigadores en Psicología del Mercosur. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires, 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.

Aguerri, María Ester, Picón Janeiro, Jimena, Abal, Facundo Juan Pablo y Galibert, María Silvia.

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

Aguerri, María Ester; Picón Janeiro, Jimena; Abal, Facundo Juan Pablo; Galibert, María Silvia
Facultad de Psicología, Universidad de Buenos Aires, UBACyT

RESUMEN
En el presente estudio de simulación se analiza la posible incidencia de la longitud del test en la detección errónea del funcionamiento diferencial del ítem (Differential Item Functioning, DIF), de las pruebas de Breslow-Day y reglas combinadas, en presencia de impacto. Las pruebas de Breslow-Day, global y de la tendencia, son aptas para detectar la presencia de DIF No Uniforme. Mientras que las reglas que combinan a las pruebas de Breslow-Day con el procedimiento estándar de Mantel-Haenszel, sin el ajuste el ajuste Bonferroni, y la regla de decisión combinada (RDC) propuesta por Randall Penfield son aplicables a la detección de cualquier tipo de DIF. Se consideraron las respuestas simuladas, sin DIF y en presencia de impacto, a tests de 40 y de 75 ítems. En todos los casos, el tamaño muestral de los grupos de Referencia y Focal fue 1,000. Realizadas 100 repeticiones se obtuvo la proporción de detección errónea del DIF. La longitud del test no resultó influyente para las pruebas de Breslow-Day al 5% pues presentaron tasas de detección errónea próximas a la nominal en las dos situaciones; sí resultó influyente para RDC, que al 5% mostró resultados satisfactorios sólo para tests largos.

Palabras clave
DIF Breslow-Day Mantel-Haenszel

ABSTRACT
EFFECT OF THE TEST LENGTH ON THE ERRONEOUS DETECTION OF THE DIFFERENTIAL ITEM FUNCTIONING OF THE BRESLOW-DAY TESTS AND COMBINED RULES. A STUDY IN THE PRESENCE OF IMPACT
This simulation study sets out to analyze the possible effect of the test length on the erroneous detection of the differential item functioning (DIF) of the Breslow-Day tests and combined rules in the presence of impact. The Breslow-Day tests, both global and of trend, are suitable to detect Nonuniform DIF. The decision rules that combine each one of them with the Mantel-Haenszel procedure (MH), without Bonferroni's adjustment, and the Combined Decision Rule (CDR), put forward by Randall Penfield, are applicable to the detection of any type of DIF. Simulated responses to 40 and 75-item tests, without DIF and with the presence of impact, were considered. The sample size of the Reference and Focal group was 1,000 in all the cases. According to the results obtained after 100 repetitions, no effect of the test length was observed in the erroneous DIF detection rate of the Breslow-Day tests at 5%, since they presented rates that neared the nominal value in both situations. However, the test length did affect the CDR at 5%, yielding satisfactory results only in the case of long tests.

Key words
DIF Breslow-Day Mantel-Haenszel

La evaluación del funcionamiento diferencial del ítem (*Differential Item Functioning*, DIF) aporta información importante a los estudios sobre la validez de los instrumentos de medición. Se dice que un ítem presenta DIF cuando sujetos de distintos grupos, pero de un mismo nivel de habilidad en el rasgo medido, tienen diferente posibilidad de responderlo correctamente. La presencia de DIF indica que una variable extraña a la habilidad que se pretende medir influye en la respuesta al ítem y por tanto pone en evidencia la falta de validez del instrumento. Camilli y Shepard (1994) mencionan los llamados métodos de Tablas de Contingencia para detectar DIF. Tales métodos se basan en el análisis de las tablas 2x2 que resultan de considerar los grupos que intervienen (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 y la prueba global de Breslow-Day para la heterogeneidad de los OR son mencionadas por Camilli y Shepard (1994) para detectar, respectivamente, la presencia de DIF y de DIF No Uniforme. La longitud del test, la diferencia de los grupos en cuanto a la habilidad medida o al tamaño muestral, pueden conducir, entre otros factores, a detecciones erróneas del DIF. Acerca de la prueba de Mantel-Haenszel (MH) es mucho lo que se ha estudiado, pero no ocurre lo mismo con la prueba global de Breslow-Day ni con la prueba de Breslow-Day de la tendencia en la heterogeneidad de los OR. El objetivo del presente trabajo es analizar la posible incidencia de la longitud del test en la detección errónea del DIF de las pruebas de Breslow-Day y reglas combinadas cuando los grupos intervinientes difieren en la habilidad medida, es decir, en presencia de impacto.

PROCEDIMIENTOS DE DETECCIÓN DEL DIF

Pruebas de Breslow-Day

Breslow y Day (1980) presentaron pruebas estadísticas para decidir 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 presentadas en el marco de estudios sobre cáncer y son profusamente utilizadas en el campo de la epidemiología. Dos de ellas son aplicables a la detección del DIF No Uniforme. En Aguerri, Galibert, Lozzia y Attorresi (2004) y Aguerri, Galibert, Attorresi y Prieto-Marañón (en prensa) se aplicó la prueba global de Breslow-Day (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 Odds Ratio (BDT) sobre un test de 40 ítems. Por otra parte, en Aguerri, Galibert, Attorresi y Prieto-Marañón (2007) se aplicaron ambas pruebas de Breslow-Day sobre un test de 75 ítems.

Procedimientos de Mantel-Haenszel

El procedimiento MH, originalmente presentado por Mantel y Haenszel (1959) en el marco de estudios sobre 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 por el bajo costo computacional y su sencillez conceptual. 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. Un análisis sobre las muestras completas y otros dos sobre las submuestras que resultan de considerar los sujetos, de ambos grupos, con puntuaciones que no superan a la media del grupo total, y los sujetos, de ambos grupos, cuyo puntaje sí supera a la media del grupo total. Se decide que un ítem presenta DIF si así fue señalado en alguno de los tres análisis. 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.

Reglas combinadas

En este trabajo se han empleado reglas para la detección del DIF, cada una de las cuales se basa en la decisión de dos pruebas de hipótesis realizadas ambas con nivel de significación α. Estas 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 también se ha utilizado la Regla de Decisión Combinada (RDC) propuesta por Penfield (2003). Esta regla consiste en decidir que un ítem presenta DIF si MH o BDT lo detectan, cada una de estas pruebas con nivel α/2. Esta 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 cccrossing SIBTEST en cuanto al error de Tipo I. Mientras que en Aguerri et al. (2007) se mostró que RDC tiene, al 5%, una tasa de Error de Tipo I semejante a la esperada.

MÉTODO

Se realizó un estudio de simulación para evaluar la proporción de DIF erróneamente detectado con BD, BDT, MH, MHoBD, MHoBDT, MHmo y RDC sobre un diseño similar al de Aguerri et al. (2007), exceptuada la longitud del test. Las respuestas a tests de 40 ítems (longitud moderada) 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 discriminación y dificultad de 20 ítems fueron prefijados y, para mayor realismo, los parámetros de los ítems restantes se tomaron de una calibración sobre datos reales (Fidalgo, Mellenbergh & Muñiz, 1999). En todos los casos los grupos, de Referencia y Focal, son de tamaño 1,000. La situación de impacto se introdujo considerando que los sujetos del grupo de Referencia pertenecen a una población cuya habilidad se distribuye como una normal estándar mientras que los sujetos del grupo Focal pertenecen a una población normal con habilidad media -1 y desvío 1. 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 los métodos mencionados 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 40 ítems, cuando se trabajó al 1% fue 0.0142 para BD, 0.0185 para BDT, 0.0397 para MH, 0.0535 para MHoBD, 0.0578 para MHoBDT, 0.0675 para MHmo y 0.0362 para RDC. Las proporciones correspondientes cuando se trabajó al 5% fueron: 0.0655, 0.0677, 0.1177, 0.1745, 0.1767, 0.2027 y 0.1092. Según Aguerri et al. (2007) sobre tests de 75 ítems tales proporciones fueron: 0.018 para BD, 0.0137 para BDT, 0.0188 para MH, 0.0365 para MHoBD, 0.0324 para MHoBDT, 0.0388 para MHmo y 0.018 para RDC al 1%; y 0.0695, 0.0605, 0.0717, 0.1359, 0.1273, 0.1467 y 0.0717 al 5%. Por tanto, ningún método verificó la condición estricta de Bradley (1978), esto es, elegido un nivel de significación α, que la proporción de detección errónea resulte comprendida entre 0.9*α y 1.1*α. Los métodos que verificaron la condición liberal de Bradley, proporción de detección errónea comprendida entre 0.5*α y 1.5*α al 5% fueron BD y BDT tanto para los tests de 40 ítems como para los de 75 ítems, y MH y RDC para los tests largos; al 1% la condición liberal fue verificada sólo por BD para los tests de longitud moderada y por BDT para los largos.

DISCUSIÓN

La longitud del test no resultó influyente en la tasa de detección errónea del DIF para las pruebas de Breslow-Day al 5% pues tanto para tests de longitud moderada como largos se verificó la condición liberal de Bradley. La prueba global de Breslow-Day verificó la condición liberal al 1% para los tests de 40 ítems y la excedió levemente para los tests de 75 ítems mientras que la prueba de Breslow-Day para la tendencia verificó la situación inversa. La longitud del test sí resultó influyente en la tasa de detección errónea para la prueba de Mantel-Haenszel, y en consecuencia también lo fue para las reglas combinadas. La prueba de Mantel-Haenszel y la Regla de Decisión Combinada propuesta por Penfield verificaron la condición liberal de Bradley al 5% únicamente para tests de 75 ítems. Las reglas combinadas sin el ajuste de Bonferroni presentaron una tasa de falsos positivos inflada, particularmente cuando se trabajó al 1%, tanto para los tests de 40 como para los de 75 ítems. Sin embargo fueron menos infladas que las del procedimiento de Mantel-Haenszel modificado.

Por tanto, si se estudia el DIF en tests de longitud moderada con presencia de impacto ha de tenerse presente que sólo a las pruebas de Breslow-Day al 5% les correspondió una tasa de detección errónea semejante al valor nominal. En tests largos, y en cuanto al riesgo de detección errónea, ha de preferirse la Regla de Decisión Combinada de Penfield al 5% pues presentó una tasa de falsos positivos próxima a la nominal y es apta para detectar ambos tipos de DIF.

En base a futuros estudios sobre la potencia de las pruebas de Breslow-Day y las reglas combinadas podrán establecerse criterios para decidir, en condiciones similares, cuál es el procedimiento más recomendable para estudiar el DIF según la longitud del test.

BIBLIOGRAFÍA

AGUERRI, M.E.; GALIBERT, M.S.; ATTORRESI, H.F. & PRIETO-MARAÑÓN, P. (2007, noviembre). Detección errónea del Funcionamiento Diferencial del Ítem. Una comparación en presencia de impacto de las pruebas de Breslow-Day, los procedimientos de Mantel-Haenszel y reglas combinadas. Ponencia libre presentada en el III Congreso Marplatense de Psicología, Mar del Plata, Argentina.

AGUERRI, M.E.; GALIBERT, M.S.; ATTORRESI, H.F. & PRIETO-MARAÑÓN, P. (en prensa). Erroneous Detection of Nonuniform DIF using the Breslow-Day test in a short test. Quality and Quantity. International Journal of Methodology.

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.

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.

FIDALGO, A.M.; MELLENBERGH, G. . & MUÑIZ, J. (1999). Aplicación en una etapa, dos etapas e iterativamente de los estadísticos Mantel-Haenszel. Psicológica, 20, 227-242.

HOLLAND, P.W. & THAYER, D.T. (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.