XXVII Congreso de la Asociación Latinoamericana de Sociología. VIII Jornadas de Sociología de la Universidad de Buenos Aires. Asociación Latinoamericana de Sociología, Buenos Aires, 2009.

Detección errónea de las pruebas de breslow-day y reglas combinadas aplicadas al análisis. Del funcionamiento diferencial del ítem en un test corto. Un estudio sobre muestras pequeñas.

María Ester Aguerri, Jimena Picón-Janeiro, Germán Diego Blum, Gabriela Susana Lozzia, Facundo Juan Pablo Abal, María Silvia Galibert y Horacio Félix Attorresi.

Detección errónea de las pruebas de breslow-day y reglas combinadas aplicadas al análisis

Del funcionamiento diferencial del ítem en un test corto Un estudio sobre muestras pequeñas

María Ester Aguerri
aguerri@psi.uba.ar

Jimena Picón-Janeiro
jpicon@psi.uba.ar

Germán Diego Blum
blumworx@gmail.com

Gabriela Susana Lozzia
glozzia@psi.uba.ar

Facundo Juan Pablo Abal
fabal@psi.uba.ar

María Silvia Galibert
galibert@psi.uba.ar

Horacio Félix Attorresi
horacioattorresi@fibertel.com.ar

Instituto de Investigaciones, Facultad de Psicología,
Universidad de Buenos Aires, Argentina[1].

[1] La investigación que se presenta en este artículo fue realizada con subsidios de la Universidad de Buenos Aires (UBACyT P043) y de la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT PICT 20909).

En los estudios actualizados sobre la validez de los instrumentos de medición se recomienda realizar el análisis del funcionamiento diferencial del ítem (*Differential Item Functioning, DIF*). Un ítem presenta DIF cuando sujetos de distintos grupos y de un mismo puntaje total en el test, no tienen la misma posibilidad, o chance, de respuesta correcta. El DIF así caracterizado, prescindiendo de cualquier modelo de medición, corresponde a los llamados métodos de Tablas de Contingencia. La presencia de DIF afecta los resultados de la medición artificialmente, pues sujetos de un mismo puntaje total, es decir del mismo nivel en la habilidad medida, se ven favorecidos, o perjudicados, según el grupo de pertenencia. Para cada grupo y nivel del puntaje total, la posibilidad de respuesta correcta al ítem (*Odds*) se obtiene como el cociente entre la cantidad de sujetos del grupo que lo respondió correctamente y la cantidad de sujetos del mismo grupo que lo respondió incorrectamente. Por lo general, y es el caso de este trabajo, se consideran dos grupos, identificados como grupo de Referencia (GR) y grupo Focal (GF). También puede estudiarse el DIF cuando hay más de dos grupos, Penfield (2001) presenta diversos procedimientos para su análisis. Por otra parte, en este trabajo nos circunscribiremos al estudio del DIF de ítems dicotómicos, o bien, aquellos cuyas posibles respuestas hayan sido dicotomizadas. Para los ítems politómicos también puede realizarse el análisis del DIF, entre otros autores, Zwick y Thayer (1996), Zwick, Thayer y Mazzeo (1997) presentan métodos para su abordaje.

Conocido el cociente de las posibilidades de GR y GF (*Odds Ratio, OR*) para todos los niveles del puntaje total, puede señalarse la presencia de DIF y clasificarlo. Si para todo nivel del puntaje total, el OR: a) es igual a 1, el ítem no tiene DIF, pues dentro de cada nivel de habilidad los sujetos de ambos grupos tienen la misma posibilidad de responderlo correctamente, b) es constante y distinto de 1, hay DIF Uniforme, y c) no es constante, hay DIF no Uniforme. Camilli y Shepard (1994) presentan distintos procedimientos para la detección del DIF de ítems dicotómicos respondidos por dos grupos. Mencionan, dentro del conjunto de los métodos de Tablas de Contingencia, a la prueba de Mantel-Haenszel (MH) y a la prueba global de Breslow-Day para la heterogeneidad de los OR (BD) para decidir, respectivamente, acerca de la presencia de DIF y de DIF no Uniforme.

La prueba MH, originalmente propuesta por Mantel y Haenszel (1959) en el marco de estudios epidemiológicos 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 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 *a*. 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.

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, como MH, para ser empleadas en estudios sobre el cáncer, y son aplicables a la detección del DIF no Uniforme. Una es la prueba global de Breslow-Day (BD) y la otra es la prueba de la tendencia en la heterogeneidad de los OR (BDT).

Acerca del procedimiento MH es mucho lo que se ha estudiado, pero no ocurre lo mismo con las pruebas de Breslow-Day. Aguerri, Galibert, Lozzia y Attorresi (2004) y Aguerri, Galibert, Attorresi y Prieto-Marañón (2009) aplicaron BD al estudio del DIF sobre un test de 20 ítems. Penfield (2003) y Picón-Janeiro et al. (2008, Noviembre) utilizaron, entre otros métodos, a BDT sobre un test de 40 ítems. En Aguerri et al. (2008, Julio) se analizó la tasa de falsos positivos de BDT y BD en tests de 20, 40 y 75 ítems, libres de DIF, respondidos por muestras de 1,000 sujetos.

La prueba MH tiene alta potencia para la detección del DIF Uniforme y las pruebas de Breslow-Day son aptas para identificar al DIF no Uniforme, por tanto se espera que las reglas que las combinen sean capaces de detectar ambos tipos de DIF. Basado en estas consideraciones, Penfield (2003) propuso la Regla de Decisión Combinada (RDC) según la cual un ítem presenta DIF, con nivel de significación *a,* si MH o BDT lo detectan, cada una de estas pruebas con nivel *a*/2. Esta modificación en el nivel de significación se denomina *ajuste de Bonferroni.* En ese mismo trabajo Penfield mostró que RDC tiene resultados, en cuanto a la tasa de falsos positivos, superiores a los de la regresión logística y del crossing SIBTEST. En el presente estudio se aplican otras dos reglas que combinan a cada una de las pruebas de Breslow-Day con MH. 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, en ambos casos las dos pruebas involucradas en la regla se realizan con nivel de significación *a*. Estas reglas no contemplan el *ajuste de Bonferroni* en el nivel de significación, como tampoco lo considera el procedimiento de Mantel-Haenszel modificado, con la ventaja de requerir dos estudios del DIF en lugar de tres.

El objetivo del presente trabajo es analizar la tasa de detección errónea, falsos positivos, de las pruebas de Breslow-Day y de las reglas que las combinan con la prueba estándar de Mantel-Haenszel (MH), incluida la regla de decisión combinada de Penfield, cuando el test es corto y las muestras son pequeñas.

Método

Se consideraron las respuestas a un test de 20 ítems de muestras de tamaño 200. Las respuestas fueron simuladas según el modelo logístico de tres parámetros[2] mediante el programa PARDSIM® (Yoes, 1997). Los parámetros de los ítems 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). Para todos los ítems el parámetro de aciertos por azar se fijó en 0.20. En todos los casos los grupos fueron elegidos de una población normal estándar. 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 y Conclusiones

La proporción de falsos positivos para los tests de 20 ítems, cuando se trabajó al 1% fue .0115 para BD y BDT, .0065 para MH, .0180 para MHoBD y MHoBDT, .0175 para MHmo y 0.0100 para RDC. Tales proporciones cuando se trabajó al 5% fueron: .0595, .0615, .0495, .1060, .1075, .1215 y .0490. Tanto las prueba de Breslow-Day como MH verificaron al menos la condición liberal establecida por Bradley (1978) al 1% y al 5%, pues la proporción de detección errónea resultó comprendida entre 0.5**a* y 1.5**a*. En particular MH al 5% verificó además la condición estricta de Bradley pues la proporción de falsos positivos resultó comprendida entre 0.9**a* y 1.1**a*. Las reglas combinadas sin el ajuste de Bonferroni, MHmo incluida, exhibieron tasas de error de Tipo I infladas. Mientras que RDC satisfizo la condición estricta de Bradley tanto al 1% como al 5%.

[2] El modelo logístico de tres parámetros permite expresar la probabilidad de respuesta a un ítem en función de la habilidad del sujeto y de tres valores fijos: el parámetro de discriminación, el parámetro de dificultad y el parámetro de aciertos por azar.

La bondad de nuevos métodos de detección del DIF se evalúa de manera comparativa, sobre un mismo diseño, con otros de eficacia reconocida. En este trabajo se compararon, en cuanto a las tasas de error de Tipo I, a las pruebas de Breslow-Day y las reglas combinadas con los procedimientos de Mantel-Haenszel, estándar y modificado. En tal sentido pudo apreciarse que las pruebas de Breslow-Day presentaron tasas de falsos positivos próximas a la nominal, al igual que el procedimiento estándar de Mantel-Haenszel. La regla de decisión combinada de Penfield resultó mejor que las reglas combinadas sin el ajuste de Bonferroni, y en particular, resultó superior al procedimiento de Mantel-Haenszel modificado. Similares resultados fueron obtenidos para tests cortos y de longitud moderada respondidos por muestras de tamaño 1,000 según Aguerri et al. (2008, Julio) y para un test de longitud moderada respondido por muestras pequeñas según Picón-Janeiro et al. (2008, Noviembre).

Será necesario realizar estudios sobre la potencia de estos procedimientos para evaluar la capacidad de detectar ítems que sí tienen DIF. Mientras tanto para el análisis del DIF sobre datos reales, en condiciones semejantes a las de este estudio, se recomienda aplicar la regla de decisión combinada de Penfield pues mantiene la tasa de detección errónea próxima a los valores nominales y es apta para detectar todo tipo de DIF.

Referencias

- 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*, 35-44.
- 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., 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. 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. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*, 235-259.
- 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.
- Picón-Janeiro, J. C., Aguerri, M. E., Blum, G. D., Lozzia, G. S., Abal, F. J. P., Galibert, M. S. & Attorresi, H. F.(2008, Noviembre). Incidencia del impacto en la detección errónea del funcionamiento diferencial del ítem de las pruebas de Breslow-Day y reglas combinadas. Un estudio para tests de longitud moderada sobre muestras pequeñas. Trabajo libre. *I Encuentro de Docentes e Investigadores Estadística en Psicología***.** Facultad de Psicología, UBA. Buenos Aires, Argentina. Del 6 al 8 de noviembre de 2008. Actas en CD pp.148-150.
- 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.
- Zwick, R. & Thayer, D. T. (1996). Evaluating the magnitude of Differential Item Functioning in polytomous items. *Journal of Educational Statistics, 21,* 187-201.
- Zwick, R., Thayer, D. T. & Mazzeo, J. (1997). Descriptive and Inferential Procedures for Assessing Differential Item Functioning in Polytomous Items. *Applied Measurement in Education, 10*, 321-344.