

III Congreso Internacional de Investigación y Práctica Profesional en Psicología
XVIII Jornadas de Investigación Séptimo Encuentro de Investigadores en
Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos
Aires, Buenos Aires, 2011.

Error de tipo I y potencia de las Pruebas de Breslow-Day y reglas combinadas aplicadas a la detección del funcionamiento diferencial del ítem no uniforme simétrico. Un estudio sobre un test largo con muestras grandes.

Blum, G. Diego, Aguerri, María Ester y Lozzia, Gabriela.

Cita:

Blum, G. Diego, Aguerri, María Ester y Lozzia, Gabriela (2011). *Error de tipo I y potencia de las Pruebas de Breslow-Day y reglas combinadas aplicadas a la detección del funcionamiento diferencial del ítem no uniforme simétrico. Un estudio sobre un test largo con muestras grandes. III Congreso Internacional de Investigación y Práctica Profesional en Psicología XVIII Jornadas de Investigación Séptimo Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires.*

Dirección estable: <https://www.aacademica.org/000-052/943>

ARK: <https://n2t.net/ark:/13683/eRwr/77D>

ERROR DE TIPO I Y POTENCIA DE LAS PRUEBAS DE BRESLOW-DAY Y REGLAS COMBINADAS APLICADAS A LA DETECCIÓN DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM NO UNIFORME SIMÉTRICO. UN ESTUDIO SOBRE UN TEST LARGO CON MUESTRAS GRANDES

Blum, G. Diego; Aguerri, María Ester; Lozzia, Gabriela

Instituto de Investigaciones, Facultad de Psicología, Universidad de Buenos Aires

RESUMEN

En este trabajo se analizaron métodos de detección del funcionamiento diferencial del ítem (DIF) en presencia de DIF no Uniforme Simétrico. Se consideraron las pruebas de Breslow-Day, global (BD) y de la tendencia (BDT) y reglas que las combinan con el procedimiento Mantel-Haenszel estándar (MH), mencionadas como MHoBD y MHoBDT. Se efectuaron comparaciones con el procedimiento de Mantel-Haenszel modificado (MHmo) y con la Regla de Decisión Combinada (RDC) de Penfield. Se consideraron las respuestas, de muestras de tamaño 1,000, a un test de 75 ítems. La presencia de DIF no Uniforme Simétrico se generó en un ítem de dificultad media. Se analizaron cuatro situaciones, sobre la base de 100 repeticiones, de manera que el área sin signo de Raju fuera 0.80. Sólo las tasas de falsos positivos (FP) de BD, BDT, MH y RDC resultan próximas al nivel nominal. Cuando la discriminación del ítem es baja todos los métodos resultan menos potentes. Cuando dicha discriminación aumenta, MH es el único que mantiene tasas de identificaciones correctas (IC) bajas mientras que BDT, MHoBDT, MHmo y RDC alcanzan prácticamente la máxima potencia. Los métodos BDT y RDC muestran la mejor relación entre las tasas de FP e IC, resultando los más recomendables.

Palabras clave

DIF no uniforme simétrico Pruebas de Breslow-Day Procedimiento Mantel-Haenszel Reglas combinadas

ABSTRACT

TYPE I ERROR AND POWER OF THE BRESLOW-DAY TESTS AND COMBINED RULES APPLIED TO THE DETECTION OF SYMMETRIC NON UNIFORM DIFFERENTIAL ITEM FUNCTIONING. A STUDY CONDUCTED ON A LONG TEST WITH LARGE SAMPLE SIZES

This study analyzed methods for detecting differential item functioning (DIF) in the presence of Symmetric non Uniform DIF. The tests included in the study were Breslow-Day Tests, both global (BD) and of trend (BDT), and rules that combine them with the standard Mantel-Haenszel (MH) procedure, referred to as MHoBD and MHoBDT. Comparisons were made with the modified Mantel-

Haenszel procedure (MHmo) and the Combined Decision Rule (CDR) of Penfield. Responses from sample sizes of 1,000 to a test of 75 items were studied. The presence of Symmetric non Uniform DIF was generated in an item of medium difficulty. Four situations were analyzed on the basis of 100 replications, so that the unsigned Raju's area was 0.80. Only the false positive (FP) rates of BD, BDT, MH and CDR are near the nominal level. When the item discrimination is low, all methods are less powerful. When such discrimination increases, MH is the only method keeping correct identification (CI) rates low whereas BDT, MHoBDT, MHmo and CDR reach almost their highest power. BDT and CDR show the best relationship between FP and CI rates, resulting in the most recommendable methods.

Key words

Symmetric non uniform dif Breslow-Day tests Mantel-Haenszel procedure Combined rules

Se conoce como Funcionamiento Diferencial del Ítem (*Differential Item Functioning*, DIF) al término estadístico empleado 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). Los métodos de Tablas de Contingencia (Camilli & Shepard, 1994) permiten evaluar la presencia de DIF mediante el análisis de tablas de 2x2, cuando se trata de dos grupos y el ítem es dicotómico. Para cada nivel del puntaje total, tales tablas resultan de cruzar el grupo de pertenencia, de Referencia (GR) o Focal (GF) según la bibliografía específica, y la respuesta al ítem, correcta o incorrecta. Para los métodos de Tablas de Contingencia las personas igualmente capaces son las que tienen igual puntaje en el test. Un ítem no presenta DIF cuando para todo nivel del puntaje total, el GF y el GR poseen la misma posibilidad (Odds) de respuesta correcta, o en otros términos, cuando el cociente de las posibilidades (Odds Ratio, OR) es 1. Si para todos los niveles del puntaje total el OR es constante y diferente de 1, el ítem posee DIF Uniforme. Si no es constante, es decir hay hetero-

geneidad de los OR, el DIF es no Uniforme.

Las pruebas de Breslow y Day (1980) permiten concluir acerca de la heterogeneidad de los OR, mientras que el procedimiento MH (Mantel & Haenszel, 1959) es el más potente para la detección del DIF Uniforme. Mazor, Clauser y Hambleton (1994) propusieron el procedimiento de Mantel-Haenszel modificado (MHmo) que resultó más potente que MH frente al DIF no Uniforme.

Penfield (2003) propuso la Regla de Decisión Combinada (RDC) según la cual un ítem presenta DIF si MH o BDT lo detectan, cada una con nivel $\alpha/2$. Tal modificación del nivel de significación se denomina *ajuste de Bonferroni*. Sobre tests de 40 ítems, Penfield estudió la tasa de error de Tipo I de RDC y su potencia frente a ítems con DIF no Uniforme Simétrico (las curvas características del ítem, CCI, sólo difieren en el parámetro de discriminación). En Aguerri, Galibert, Attorresi y Prieto-Marañón (2007) se presentaron dos reglas que combinan, sin efectuar el *ajuste de Bonferroni*, a BD y BDT con MH, identificadas respectivamente como MHoBD y MHoBDT. BD, BDT y las reglas combinadas fueron evaluadas en tests de 75 ítems con 27% de ítems con DIF Paralelo (las CCI sólo difieren en el parámetro de dificultad), entre otros trabajos, en Aguerri, Blum, Picón, Abal, Lozzia, Galibert, y Attorresi (2010) y con 27% de ítems con DIF no Uniforme Simétrico en Aguerri et al. (2008).

El objetivo de este trabajo es evaluar a las pruebas de Breslow-Day y las reglas combinadas frente al DIF no Uniforme Simétrico cuando el ítem con DIF es de dificultad media.

Método

Se consideraron las respuestas a un test de 75 ítems. Los grupos de Referencia y Focal (GR y GF), pertenecientes a una población normal estándar, se tomaron de tamaño 1,000. Los parámetros de discriminación (a) y de dificultad (b) de 20 ítems resultaron de combinar cuatro niveles de a (0.25, 0.60, 0.90 y 1.25) con cinco niveles de b (-1.5, 1, 0, 1 1.5). Para los 55 ítems restantes tales parámetros se tomaron de una calibración sobre datos reales entre los publicados por Clauser, Mazor y Hambleton (1994). El parámetro de aciertos por azar (c) se consideró igual a 0.20 en todos los casos. Las respuestas fueron simuladas mediante PARDSIM® (Yoes, 1997) según el modelo logístico de tres parámetros. La presencia de DIF no Uniforme Simétrico se generó en un solo ítem de dificultad media ($b = 0$). Se investigaron cuatro situaciones con distintos niveles de discriminación del ítem bajo estudio en el GR (0.25, 0.6, 0.9 o 1.25). Los valores correspondientes en el GF se tomaron de manera que el área sin signo de Raju (1988) fuera aproximadamente igual a 0.80 (0.36, 2.25, 0.43 y 0.495 respectivamente). Para cada situación se simuló 100 pares de patrones de respuesta. Mediante el Programa Computacional Bday (Prieto-Marañón, 2005) se estudió el DIF con BD, BDT, MH, MHoBD, MHoBDT, MHmo y RDC al 5%. El programa Bday aplica el proce-

dimiento MH, como el PROC FREQ de SAS (Statistical Analysis System, 1989), en una sola etapa y sin la corrección por continuidad. Finalmente se registraron las tasas de falsos positivos (FP) y de identificaciones correctas (IC) para cada uno de los métodos bajo estudio. Se analizó si los métodos satisfacen, o no, el criterio estricto o el liberal de Bradley (1978). Para satisfacer la condición estricta, la tasa de FP debe estar comprendida entre .045 y .055 mientras que para satisfacer la liberal, los valores extremos son .025 y .075.

Resultados y Discusión

La Tabla 1 contiene las proporciones de FP en el test y de IC al 5% según el método y el parámetro de discriminación del ítem en GR.

Las tasas de FP de MHoBD, MHoBDT y MHmo duplican el valor nominal para todos los niveles de discriminación del ítem con DIF, mientras que las de BD, BDT, MH y RDC satisfacen al menos la condición liberal de Bradley. En cuanto a las tasas de IC, se observa que todos los métodos presentan los valores más bajos cuando $a_{GR} = 0.25$. En el resto de los casos dichas tasas se elevan considerablemente, excepto la de MH que se mantiene muy baja. Como consecuencia existe un importante contraste entre la tasa media de IC de MH y las del resto de los métodos. El contraste máximo se da cuando $a_{GR} = 0.60$, donde todos los métodos excepto MH detectan el 100% de las veces al ítem con DIF. Cuando $a_{GR} = 0.90$, las tasas de IC de BD y MHoBD son relativamente altas, pero no tan altas como cuando $a_{GR} = 0.60$ o $a_{GR} = 1.25$. Las tasas de IC de BDT, MHoBDT, MHmo y RDC son próximas a 1 en tres niveles de a_{GR} (0.60, 0.90 y 1.25). De estos métodos, los que exhiben la relación entre tasas de FP y de IC más apropiada son BDT y RDC, puesto que cumplen al menos la condición liberal de Bradley.

Tabla 1
Proporción de Falsos Positivos en el Test y de Identificaciones Correctas del ítem con DIF no Uniforme Simétrico al 5% según el Método y el Parámetro de Discriminación en el Grupo de Referencia

Método	Tasa de Falsos Positivos en el Test				Tasa de Identificaciones Correctas			
	Parámetro de Discriminación del ítem				Parámetro de Discriminación del ítem			
	0.25	0.60	0.90	1.25	0.25	0.60	0.90	1.25
BD	.0703*	.0651*	.0654*	.0723*	.15	1	.56	.87
BDT	.0557*	.0591*	.0592*	.0516**	.37	1	.99	1
MH	.0535**	.0501**	.0526**	.0549**	.03	.04	.08	.07
MHoBD	.1209	.1112	.1145	.122	.18	1	.61	.87
MHoBDT	.1055	.1057	.1073	.1031	.40	1	.99	1
MHmo	.1203	.1189	.1139	.1201	.29	1	.98	.99
RDC	.0541**	.057*	.0538**	.0532**	.22	1	.99	1

Nota. * indica que se verifica la condición liberal de Bradley; ** indica que se verifica la condición estricta de Bradley

BIBLIOGRAFÍA

Aguerri, M.E., Blum, G.D., Picón Janeiro, J., Abal, F.J.P., Lozzia, G.S., Galibert, M.S., & Attorresi, H.F. (2010). Reglas de detección del Funcionamiento Diferencial del Ítem. Estudio del efecto del Impacto. Memorias del II Encuentro de Docentes e Investigadores de Estadística en Psicología. Facultad de Psicología, UBA. Buenos Aires. Libro de memorias, Tomo II (ID-105). pp. 5-7.

Aguerri, M. E., Galibert, M. S., Attorresi, H., & Prieto-Marañón, P. (2007). Aplicación de las pruebas de Breslow-Day a la detección del DIF. Comparación con el procedimiento de Mantel-Haenszel modificado. Presentación realizada en el Symposium: Funcionamiento Diferencial del Ítem. X Congreso de Metodología de Ciencias Sociales y de la Salud. Barcelona, España. Libro de Resúmenes, p. 83.

Aguerri, M. E., Galibert, M. S., Attorresi, H. F., & Prieto-Marañón, P. (2008). Application of the Breslow-Day tests and combined rules to the DIF detection. Trabajo Libre. III European Congress of Methodology. European Association of Methodology. Oviedo, España. Libro de Resúmenes, p. 50.

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.

Clauser, B.E., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. Journal of Educational Measurement, 31, 67-78.

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.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 284-291.

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.