II Congreso Internacional de Investigación y Práctica Profesional en Psicología XVII Jornadas de Investigación Sexto Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires, 2010.

Reglas de detección del funcionamiento diferencial del ítem. Estudio del efecto del tamaño de muestra en presencia de DIF paralelo.

Aguerri, María Ester, Blum, G. Diego, Picón Janeiro, Jimena y Galibert, María Silvia.

on student ability and test efficiency for computerized adaptive tests. Applied Psychological Measurement,16, 33-40.

LUNZ, M. A. & BERGSTROM, B. A. (1994). An empirical study of computerized adaptive test administration conditions. Journal of Educational Measurement, 31, 251-263.

MUÑIZ, J. (2005). La validez desde una óptica psicométrica. Acta Comportamentalia, 13, 9-20.

NUNNALLY, J. C. & BERNSTEIN, I. J. (1995). Teoría Psicométrica. Nueva York: McGraw-Hill.

OLEA, J. & PONSODA, V. (2003). Tests adaptativos informatizados. Madrid: UNED.

OLEA, J., REVUELTA, J., XIMÉNEZ, M. C. & ABAD, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. Psicológica, 21, 157-173.

PONSODA, V., OLEA, J., RODRIGUEZ, M.S. & REVUELTA, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. Applied Measurement in Education,12, 167 -184.

REVUELTA, J., XIMÉNEZ, M. C. & OLEA, J. (2003). Psychometric and Psychological effects of item selection and review on Computerized Testing. Educational and Psychological Measurement, 63, 791-808

STOCKING, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. Applied Psychological Measurernent, 21, 129-142.

VAN DER LINDEN, W. J. & GLAS, C. A. W. (Eds.) (2000). Computerized adaptive testing. Theory and practice. Dordrecht: Kluwer Academic Publishers.

VISPOEL, W. P. (1998). Reviewing and changing answers on computer-adaptive and selfadaptive vocabulary tests. Journal of Educational Measurernent, 35, 328-345.

VISPOEL, W. P., ROCKLIN, T. R. & WANG, T. (1994). Individual differences and test administration procedures. A comparison of fixed-item, computerized-adaptive and self-adapted testing. Applied Psychological Measurernent, 7, 53-79

WISE, S. L. (1999). Tests autoadaptados informatizados: Fundamentos, resultados de investigación e implicaciones para la práctica. En J. Olea, V. Ponsoda & G. Prieto (Eds.). Tests informatizados: Fundamentos y aplicaciones. (pp. 189-206). Madrid: Pirámide.

WISE, S. L., FREEMAN, S. A., FINNEY, S. J., ENDERS, C. K. & SEVERANCE, D. D. (1997). The accuracy of examinee judgments of relative item difficulty: Implications for computerized adaptive testing. Annual Meeting of National Council of Measurement in Education. Chicago, IL.

REGLAS DE DETECCIÓN DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM. ESTUDIO DEL EFECTO DEL TAMAÑO DE MUESTRA EN PRESENCIA DE DIF PARALELO

Aguerri, María Ester; Blum, G. Diego; Picón Janeiro, Jimena; Galibert, María Silvia
UBACyT, Instituto de Investigaciones, Facultad de Psicología, Universidad de Buenos Aires

RESUMEN

En el presente estudio se analiza la posible incidencia del tamaño de muestra en la detección del funcionamiento diferencial del ítem (Differential Item Functioning, DIF) frente al DIF Paralelo. Se analizó la tasa de falsos positivos e identificaciones correctas de dos reglas que combinan a las pruebas de Breslow-Day, global (BD) y de la tendencia (BDT), con el procedimiento estándar de Mantel-Haenszel (MH), sin el ajuste de Bonferroni. Se efectuaron comparaciones con la regla propuesta por Randall Penfield y con el procedimiento de Mantel-Haenszel modificado. Las respuestas a tests de 75 ítems fueron simuladas, sin impacto, con 100 repeticiones. Los grupos de Referencia y Focal se tomaron de igual tamaño, 1,000 o 200. La presencia de DIF Paralelo se generó en 20 ítems con un marcado incremento del parámetro de dificultad en el grupo Focal. El procedimiento MH resultó muy afectado por el tamaño de muestra, con tasas de falsos positivos muy infladas para las muestras grandes. Efecto que se trasladó a los cuatro procedimientos. Para ambos tamaños muestrales, la regla basada en BDT sin el ajuste de Bonferroni resultó la más potente y la regla de Penfield presentó la tasa de falsos positivos menos inflada.

Palabras clave
Funcionamiento diferencial del ítem Pruebas Breslow-Day Procedimiento Mantel-Haenszel DIF paralelo

ABSTRACT

THE DIFFERENTIAL ITEM FUNCTIONING TESTING RULES. THE EFFECT OF THE SAMPLE SIZE IN PRESENCE OF PARALLEL DIF

This study analyzes the potential effect of the sample size on the differential Item functioning (DIF) detection in presence of Parallel DIF. The false positives and correct identifications rates of two rules that combine the Breslow-Day tests, global (BD) and of trend (BDT), with the standard Mantel-Haenszel (MH) procedure, without the Bonferroni's adjustement were analyzed. Comparisons were made with the rule proposed by Randall Penfield and the modified Mantel-Haenszel procedure. The responses to 75-items tests were simulated without impact with 100 repetitions. Reference and Focal groups were of equal size, 1,000 or 200. The presence of Parallel DIF was generated in 20 items with a marked increase of the difficulty parameter in the Focal group. The MH procedure was strongly affected by the sample size, with very inflated false positives rates for large samples. This effect moved to the four procedures. For both sample sizes, the rule based on BDT without the Bonferroni's adjustement was the most powerful and the Penfield's rule presented the false positives rates less inflated.

Key words
Differential item functioning Breslow-Day tests Mantel-Haenszel Procedure Parallel DIF

Los instrumentos de medición cuentan actualmente con nuevos recursos para el estudio de su validez, entre los que se destaca el análisis del Funcionamiento Diferencial del Ítem (*Differential Item Functioning*, *DIF*). Se utiliza el término estadístico DIF 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). Entre los métodos de detección del DIF están los llamados de Tablas de Contingencia (Camilli & Shepard, 1994). Estos procedimientos analizan, para cada ítem, tantas tablas de contingencia del tipo 2x2 (si se consideran las respuestas de dos grupos y el ítem es dicotómico) como niveles tenga el puntaje total en el test. Los grupos intervinientes se suelen identificar como grupo de Referencia (GR) y grupo Focal (GF) y las respuestas al ítem están categorizadas como correcta-incorrecta. Para el estudio del DIF de ítems politómicos y/o respondidos por dos o más grupos puede mencionarse, entre otros trabajos, a Fidalgo, Quintanilla, Fernández, Pons y Aguerri (2010). Para cada grupo y nivel del puntaje total, se define la posibilidad de respuesta correcta al ítem (*Odds*) como el cociente entre la cantidad de sujetos que respondió al ítem correctamente y la cantidad de sujetos que lo respondió incorrectamente. Cuando ambos grupos presentan la misma posibilidad de respuesta correcta al ítem a lo largo de todos los niveles del puntaje total, se dice que el ítem no presenta DIF; en otros términos, no existe DIF cuando el cociente de las posibilidades (*Odds Ratio, OR*) vale 1 para todo nivel del puntaje total. Si para todos los niveles del puntaje total los OR son iguales entre sí, pero diferentes de 1, el ítem presenta DIF Uniforme; si no se mantienen iguales, es decir hay heterogeneidad de los OR, el ítem presenta DIF no Uniforme. Holland y Thayer (1988) aplicaron el procedimiento de Mantel-Haenszel (MH) para detectar la presencia de DIF. Este método, 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 los niveles del puntaje total, y por su sencillez conceptual es profusamente utilizado para el análisis del DIF. Mazor, Clauser y Hambleton (1994) propusieron el procedimiento Mantel-Haenszel modificado (MHmo), que consiste en realizar tres análisis de DIF con un mismo nivel de significación α. Estos autores mostraron que MHmo es más potente que MH para detectar DIF no Uniforme. Por otro lado, Breslow y Day (1980) presentaron dos pruebas estadísticas para decidir sobre la heterogeneidad de los OR. Una es la prueba global de la homogeneidad de los OR (BD) y la otra es la prueba de la tendencia en la heterogeneidad de los OR (BDT). El rechazo de la hipótesis nula, tanto con BD como con BDT, conduce a señalar la presencia de DIF no Uniforme. Según los modelos de la Teoría de Respuesta al Ítem la probabilidad de respuesta correcta al ítem puede formularse en función de la habilidad del sujeto y de los parámetros del ítem. En particular el modelo logístico de tres parámetros permite expresar dicha probabilidad en función de la habilidad del sujeto y de tres valores fijos: el parámetro de discriminación (*a*), el parámetro de dificultad (*b*) y el parámetro de aciertos por azar (*c*). Su representación gráfica es la Curva Característica del Ítem (CCI). Cuando los parámetros del ítem son los mismos en GR y GF, es decir las CCIs son coincidentes, se dice que el ítem no tiene DIF. Si las CCIs sólo difieren en parámetro de dificultad, y el parámetro de aciertos por azar es no nulo, el ítem presenta DIF no Uniforme Paralelo. Si bien Mazor, Clauser y Hambleton (1994), entre otros autores, mencionaron esta última situación como DIF Uniforme, de los lineamientos de Hanson (1998) se infiere que el DIF presente es no Uniforme. Hanson llamó DIF Paralelo a tal tipo de DIF. Más allá de las denominaciones utilizadas, en tanto las CCIs difieran en alguno de los parámetros hay DIF. Penfield (2003) evaluó a BDT en tests de 40 ítems y propuso la Regla de Decisión Combinada (RDC) basada en BDT y MH con el *ajuste de Bonferroni* en el nivel de significación. Este ajuste consiste en considerar para cada prueba el nivel de significación estipulado, α, dividido por el número de pruebas realizadas, es decir, en este caso realizar cada prueba con nivel α/2. Aguerri, Galibert, Attorresi y Prieto-Marañón (2007, Febrero) presentaron dos reglas que combinan, sin el *ajuste de Bonferroni*, a BD y BDT con MH, identificadas como MHoBD y MHoBDT respectivamente. La tasa de falsos positivos de tales reglas combinadas fue evaluada, entre otros trabajos, en Aguerri, Picón-Janeiro, Blum, Abal, Lozzia y Galibert (2009). En Aguerri et al. (2008, Julio; 2009, Septiembre) se analizó tanto la tasa de falsos positivos como la potencia de las pruebas de Breslow-Day y de las reglas combinadas en tests largos frente a ítems que difieren moderadamente en el parámetro de dificultad y/o en el parámetro de discriminación. Tanto las reglas combinadas sin *el ajuste de Bonferroni*, como RDC y MHmo son aptos para detectar ambos tipos de DIF pues MH es considerado el más potente ante el DIF Uniforme y las pruebas de Breslow-Day son aptas para detectar el DIF no Uniforme. El objetivo del presente trabajo es continuar con la evaluación de las reglas basadas en las pruebas de Breslow-Day y el procedimiento de Mantel-Haenszel aplicadas al estudio del DIF en tests largos. En particular se estudiará la incidencia del tamaño de muestra cuando los ítems presentan DIF debido a una marcada diferencia de los parámetros de dificultad.

MÉTODO

Se simularon respuestas un test de 75 ítems mediante el programa PARDSIM® (Yoes, 1997) para muestras de igual tamaño, 1,000 y 200, según el modelo logístico de tres parámetros. Los parámetros de dificultad y discriminación de los ítems con DIF en GR 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 el GF los mismos ítems tuvieron igual parámetro de discriminación que en el GR y el respectivo parámetro de dificultad incrementado en 1. Esta es la mayor diferencia en los parámetros de dificultad considerada en Mazor et al. (1994). Para los 55 ítems libres de DIF se consideraron valores de los parámetros de discriminación y dificultad tomados de una administración real. En todos los casos el parámetro de aciertos por azar se fijó en 0.20 y los grupos fueron tomados de una población normal estándar, es decir sin presencia de impacto. Mediante el programa computacional Bday (Prieto-Marañón, 2005) se estudió el DIF con BD, BDT, MH, MHoBD, MHoBDT, MHmo y RDC. Se registró la tasa de falsos positivos y de identificaciones correctas, al 1% y al 5%, sobre 100 repeticiones. 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.

RESULTADOS Y CONCLUSIONES

Cuando las muestras son de tamaño 1,000, la tasa de falsos positivos al 5% dio .0736 para BD, .1120 para BDT, .4433 para MH, .4842 para MHoBD, .4942 para MHoBDT, .5342 para MHmo y .3847 para RDC. Para las muestras de tamaño 200, tales tasas fueron, respectivamente, .1144, .1369, .1227, .2244, .2242, .1996 y .1458. En cuanto a la potencia (tasa de identificaciones correctas) al 5% para las muestras de tamaño 1,000, los resultados obtenidos fueron .1610 para BD, .5015 para BDT, .9005 para MH, .9095 para MHoBD, .9265 para MHoBDT, .9345 para MHmo y .8930 para RDC. Tales resultados para las muestras de tamaño 200, fueron respectivamente, .1240, .4180, .5350, .5945, .6405, .6165 y .5725. Se omite presentar los resultados obtenidos al 1% por cuanto conducen a conclusiones similares. Sólo BD verificó, para las muestras grandes, la condición liberal de Bradley (1978). Una prueba estadística verifica esta condición cuando el nivel de significación empírico (tasa de falsos positivos) está comprendido entre 0.5 y 1.5 veces el nivel nominal. En el caso de considerar un nivel de significación del 5%, se verifica la condición liberal de Bradley si la tasa de falsos positivos está comprendida entre 0.025 y 0.075. MH presentó una tasa de falsos positivos muy inflada cuando las muestras son grandes (octuplica la nominal) mientras excede levemente al doble del nominal cuando las muestras son pequeñas. Similar efecto, aunque menos inflado, se observó en Aguerri et al. (2008, Julio; 2009, Septiembre) en presencia de DIF Paralelo de menor magnitud. El efecto del tamaño de muestra sobre MH se trasladó a las reglas combinadas ya que presentaron tasas de falsos positivos muy infladas cuando las muestras son grandes. Dado el *ajuste de Bonferroni*, RDC presentó para ambos tamaños muestrales, la tasa de falsos positivos menos inflada. La reducción del tamaño de muestra ocasionó una disminución en la potencia de todos los métodos empleados, siendo MH y las reglas combinadas los más afectados. La regla

más potente fue MHoBDT, prácticamente iguala a MHmo cuando las muestras son grandes y lo supera cuando son pequeñas. De los resultados obtenidos, para el estudio del DIF con datos reales, e independientemente del tamaño muestral, se recomienda la aplicación de MHoBDT si se prioriza la potencia, o bien RDC si se quiere reducir el riesgo de cometer el error de Tipo I.

BIBLIOGRAFIA

AGUERRI, M. E.; GALIBERT, M. S.; ATTORRESI, H. F., & PRIETO MARAÑÓN, P. (2007, Febrero). Aplicación de las pruebas de Breslow-Day a la detección del DIF. Comparación con el procedimiento de Mantel-Haenszel modificado. Simposio sobre el DIF. X Congreso de Metodología de Ciencias Sociales y de la Salud. Asociación Española de Metodología de Ciencias del Comportamiento. Barcelona, España. Libro de Resúmenes, p. 83.

AGUERRI, M. E.; GALIBERT, M. S.; ATTORRESI, H. F., & PRIETO MARAÑÓN, P. (2008, Julio). 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.

AGUERRI, M. E.; GALIBERT, M. S.; ATTORRESI, H. F., & PRIETO MARAÑÓN, P. (2009, Septiembre). Potencia y Error de Tipo I de Reglas que combinan a las Pruebas de Breslow-Day con el Procedimiento Estándar de Mantel-Haenszel. Un estudio sobre muestras pequeñas. Trabajo Libre. XI Congreso de Metodología de Ciencias Sociales y de la Salud. Asociación Española de Metodología de Ciencias del Comportamiento.Málaga, España. Libro de Resúmenes, p. 65.

AGUERRI, M. E., PICÓN-JANEIRO, J., BLUM, G., ABAL, F., LOZZIA, G., & GALIBERT, M. S. (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. Trabajo libre. Memorias del I Congreso Internacional de Investigación y Práctica Profesional en Psicología, XVI Jornadas de Investigación y V Encuentro de Investigadores en Psicología del MERCOSUR. Buenos Aires, Argentina. Actas en CD, 3, 472-473.

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., QUINTANILLA, L., FERNÁNDEZ, R., PONS, F., & AGUERRI, M. E. (2010) Detección del DIF en ítems politómicos mediante el uso de los métodos Mantel-Haenszel. Revista Electrónica de Metodología Aplicada, 15, 12-18. http://www.psico.uniovi.es/REMA/v15n1/indice.html

HANSON, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. Journal of Educational and Behavioral Statistics, 23, 244-253.

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.

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.

ESTUDIO EXPLORATORIO DE LAS PROPIEDADES PSICOMÉTRICAS DE LA ESCALA DE AUTOEFICACIA CREATIVA EN POBLACIÓN ARGENTINA

Aranguren, Maria; Irrazabal, Natalia
Centro de Investigaciones en Psicología y Psicopedagogía, Facultad de Psicología y Educación, Pontificia Universidad Católica. Argentina

RESUMEN

El presente estudio tiene como objetivo presentar algunos resultados preliminares de la adaptación de la Escala de Autoeficacia Creativa (EAC - Creativity Self-efficacy Scale) en población argentina. La EAC fue diseñada por Yi, Scheithauer, Lin y Schwarzer (2008) con la finalidad de obtener un instrumento breve y adecuado para la evaluación de la autoeficacia en un dominio específico, el de la creatividad. Los resultados obtenidos en este estudio, indican, de manera preliminar, su adecuado funcionamiento en nuestro medio así como también algunos interrogantes a dilucidar en futuras investigaciones.

Palabras clave
Autoeficacia Creatividad Propiedades psicométricas

ABSTRACT

EXPLORATORY STUDY OF THE PSYCHOMETRIC PROPERTIES OF THE CREATIVE SELF-EFFICACY SCALE IN ARGENTINIAN POPULATION.

The principal aim of this study is to present some preliminary results of the adaptation of the Creative Self-Efficacy Scale in Argentinian population. The CSE was originally designed by Yi, Scheithauer, Lin and Schwarzer (2008) in order to obtain a short and appropriate instrument for measuring self-efficacy in a particular domain, creativity domain. The results of this study indicate, on preliminary basis, its proper functioning in our environment as well as some questions to be elucidated in future research.

Key words
Self-efficacy Creativity Psychometric properties

INTRODUCCION

La creatividad ha sido estudiada en relación a los rasgos de personalidad, los procesos cognitivos involucrados, el producto o logro resultante y, por último, en relación a las características ambientales que favorecen o perjudican el potencial creativo. En particular, diferentes estudios han evidenciado que las personas pueden presentar determinados rasgos asociados a la creatividad pero de su sola presencia no se desprende la realización de esa capacidad en un determinado producto o logro creativo. En este sentido, algunos autores han propuesto estudiar la autoeficacia creativa como constructo mediador entre la persona y la realización del producto (e.g. Beghetto, 2006; Laws, 2003; Tierney & Farmer, 2002).

La autoeficacia creativa representa una extensión de un constructo más amplio, que es la autoeficacia (Beghetto, 2006). Esta ha sido definida como "los juicios de cada individuo sobre sus capacidades, en base a los cuales organizará y ejecutará sus actos que le permitan alcanzar el rendimiento deseado" (Bandura 1987, p. 416). Es decir que los juicios de las personas acerca de sus propias capacidades impactan sobre la motivación y el esfuerzo para realizar dichas actividades. La autoeficacia creativa consiste