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.

La evaluación mediante tests adaptativos informatizados. Experiencia subjetiva del examinado.

Abal, Facundo Juan Pablo, Lozzia, Gabriela, Aguerri, María Ester y Galibert, María Silvia.

LA EVALUACIÓN MEDIANTE TESTS ADAPTATIVOS INFORMATIZADOS. EXPERIENCIA SUBJETIVA DEL EXAMINADO

Abal, Facundo Juan Pablo; Lozzia, Gabriela; Aguerri, María Ester; Galibert, María Silvia
UBACyT - Agencia Nacional de Promoción Científica y Tecnológica - Universidad de Buenos Aires

RESUMEN
Un Test Adaptativo Informatizado (TAI) es una prueba administrada mediante una computadora donde los reactivos que se le presentan progresivamente al individuo dependen del rendimiento alcanzado con ítems previos. Este procedimiento permite obtener una medida más precisa en menos tiempo. A pesar de estas ventajas, las condiciones de aplicación de los TAIs son muy poco flexibles en comparación con la que presentan los test convencionales. Los TAIs no permiten omitir o diferir la respuesta a los ítems ni tampoco revisarlos. Esto genera un incremento significativo de ansiedad en los evaluados que pueden perjudicar el rendimiento estimado. Desde la visión de los examinados estas condiciones tornan a la evaluación frustrante e injusta, lo que amenaza a la validez aparente del test. La administración adaptativa produce una menor sensación subjetiva de éxito dado que, independientemente de la capacidad del individuo, todos tienen una similar proporción de aciertos (aproximadamente del 50%). Los intentos actuales por mejorar las condiciones de aplicación se centran principalmente en cuidar la motivación de quien responde y evitar un nivel de ansiedad que obstaculice la tarea. De nada sirve buscar una mayor confiabilidad cuando el nivel de habilidad estimado está siendo distorsionado por las experiencias subjetivas del evaluado.

Palabras clave
Test Adaptativo Informatizado TAI

ABSTRACT
EVALUATION BY COMPUTERIZED ADAPTIVE TESTS.
SUBJECTIVE EXPERIENCE OF THE EXAMINEE
A Computerized Adaptive Test (CAT) is a test administered by means of a computer where the items are given to the examinee progressively depending on the performance achieved with the previous ones. This procedure allows for more precise measurements in less time. Despite these advantages, the conditions for the application of the CATs are very little flexible in comparison with conventional tests. CATs do not permit omitting or deferring item responses nor reviewing them. This generates a significant increase of anxiety in the examinees that may affect the estimated performance. These conditions make the evaluation been perceived as frustrating and unfair from the point of view of examinees, which endangers the apparent validity of the test. The adaptive administration leads to reduced subjective sense of success because, regardless of individual's ability, everyone has a similar proportion of correct answers (approximately 50%). Current attempts to improve the conditions of the applications of CATs are mainly focused on preserve examinees' motivation and prevent a level of anxiety that may hamper the task. There is no point in looking for greater reliability when the estimated skill level is being distorted by the subjective experiences of the examinee.

Key words
Computerized Adaptive Test CAT

El avance en la informatización de las tareas de evaluación psicológica y educativa es inexorable. Para Hambleton (2004), el mayor cambio en el uso de los tests en las próximas décadas será la sustitución paulatina de la administración en formato de papel y lápiz por una práctica de evaluación con soporte computacional. Esto se debe principalmente a las ventajas psicométricas introducidas por la Teoría de Respuesta al Ítem y al avance tecnológico que posibilitó la Informática. En este contexto, el desarrollo e implementación de los Tests Adaptativos Informatizados (TAIs) ocupan un lugar privilegiado (Olea & Ponsoda, 2003; van der Linden & Glass, 2000).
Un TAI es un test administrado mediante una computadora donde los reactivos que se le presentan al individuo dependen del rendimiento alcanzado con ítems previos. Básicamente, el test se sustenta en un banco de ítems con características psicométricas conocidas y un conjunto de algoritmos que selecciona los ítems en función del nivel de habilidad que va manifestando el evaluado en cada respuesta. Si la persona contesta de manera correcta, el programa exhibirá un ítem más difícil; y si la respuesta es incorrecta, presentará un ítem más fácil. La administración de los ítems continúa hasta que se alcanza una cantidad previamente especificada o un valor prefijado de precisión o error típico (Nunnally & Bernstein, 1995).
Adaptar la dificultad del test al nivel de habilidad del evaluado conlleva enormes ventajas técnicas. Los TAIs permiten obtener una medida con un grado mayor de precisión en menos tiempo y utilizando un número inferior de ítems que los tests convencionales (Chang & Ying, 1996). Sin embargo, estas virtudes psicométricas de los TAIs se dan a expensas de condiciones de administración poco confortables para los examinados. Este aspecto origina experiencias aversivas para quienes responden un TAIs, lo que no contribuye a su aceptación social (Olea, Revuelta, Ximénez & Abad, 2000).
El objetivo de este trabajo es realizar una revisión bibliográfica de las experiencias subjetivas negativas que muestran los examinados al responder los Tests Adaptativos Informatizados. Además, explorar las soluciones provisionales que se están estudiando para encontrar condiciones de aplicación más agradables.

DESARROLLO
La calidad de la medida obtenida en la administración de un test no depende exclusivamente de las propiedades psicométricas del mismo. El estado psicológico del evaluado y su motivación son componentes críticos que condicionan la precisión y la validez de la puntuación alcanzada. Si el bienestar del individuo no ha sido el adecuado durante el proceso de evaluación, disminuyen las garantías de calidad de los datos recogidos. En las últimas décadas numerosos investigadores se interesaron por estudiar las experiencias subjetivas que perciben los individuos a medida que responden un TAI (e.g. Garrison & Baumgarten, 1986; Huff & Sireci, 2001; Lunz, Bergstrom & Wright, 1992; Stocking, 1997). La aplicación masiva de TAIs en países como Estados Unidos y Holanda permitió profundizar en los alcances y limitaciones de su uso cuando su administración tiene importantes consecuencias para quien lo responde (certificación profesional, selección de personal y evaluación educativa).
Comparada con la aplicación de tests convencionales de lápiz y papel para la medición de habilidades o rendimiento, la situación de evaluación con TAIs presenta factores estresantes adicionales. En principio, por tratarse de un tests informatizado puede despertar rechazo o mala predisposición en examinados poco familiarizados con las computadoras. No obstante, este no es un aspecto determinante dado que, en líneas generales, esta práctica es bien recibida por los evaluados (e.g. Garrison & Baumgarten, 1986, Vispoel, Rocklin & Wang, 1994). A su vez, es de esperar que la familiaridad vaya en aumento ya que el uso de las computadoras se ha tornado cada vez más habitual en la vida cotidiana.
El mayor foco de las críticas se centra en la poca flexibilidad que tienen las condiciones de aplicación de los TAIs en comparación con la que presentan los test convencionales (informatizados o no). Debe recordarse que usualmente el algoritmo de selección de los ítems de un TAI necesita conocer la respuesta al ítem previo para elegir el próximo, por lo que no es posible omitir o diferir

la repuesta. Estas condiciones borran las diferencias individuales en la manera con que se abordan los reactivos. Dadas las trascendentes consecuencias que tiene el resultado del test para el individuo, es posible considerar la situación de evaluación como estresante. El hecho de regular de forma rígida la secuencia de exposición de los ítems podría generar en el evaluado una sensación de pérdida de control sobre el test. Algunos autores suponen que esta sensación provoca en el examinado un incremento significativo de su ansiedad que puede perjudicar su rendimiento (e. g. Stocking, 1997; Wise, Freeman, Finney, Enders & Severance, 1997). Es justamente desde esta línea de investigación que se propone el desarrollo de Tests Autoadaptativos Informatizados, cuya diferencia con los TAIs radica en que los evaluados pueden elegir dentro de un determinado rango impuesto por el programa el nivel de dificultad del próximo ítem que deberán responder. Así, al percibir el control sobre el estresor se reduce la ansiedad propiciando un contexto favorable para optimizar el rendimiento del evaluado (Wise, 1999).

De todas las condiciones de aplicación cuestionadas, la más controvertida es la imposibilidad de revisar las respuestas. Algunos TAIs ni siquiera permiten modificar aquellas respuestas que fueron ingresadas incorrectamente por un accidente o por distracción. Los educadores suelen recomendar una revisión final del examen para detectar errores en el propio desempeño. Esta es una estrategia valorada y considerada también como una demostración de conocimiento legítima en la resolución de un test convencional. Más allá del incremento de ansiedad que pueda generar impedir una revisión, los examinados destacan que la prueba bajo estas condiciones es frustrante e injusta (Lunz et al. 1992); consideraciones que amenazan directamente a la validez aparente del test.

Los especialistas afirman que los evaluados sostienen fundadas razones para incluir una revisión de los ítems (Wise et al. 1997). No obstante, se suelen exponer importantes justificativos técnicos para negarla: se incrementa el tiempo de evaluación, se reduce la precisión en la estimación de la habilidad y se obtienen puntuaciones infladas artificialmente. Asimismo, argumentan aspectos vinculados a la seguridad de los datos: los examinados podrían obtener información para modificar sus respuestas en la revisión o podrían implementar estrategias ilegítimas para responder los ítems (fallar deliberadamente algún ítem para recibir ítems fáciles, y después responderlo correctamente durante la revisión) (Olea & Ponsoda, 2003; Wise et al., 1997).

Desde la perspectiva de los evaluados, la revisión de los ítems en los TAIs tiene consecuencias positivas. Los estudios de Olea et al. (2000) y Revuelta, Ximénez y Olea (2003) mostraron que no sólo disminuye la ansiedad y aumenta el confort de los examinados, sino que también se mejora significativamente el nivel de habilidad estimado sin una pérdida importante de precisión. A pesar de esto, el mejor rendimiento podría ser explicado desde dos hipótesis no antagónicas. La primera supone que, efectivamente, la revisión favorece a que el examinado experimente una reducción del estrés y así pueda cambiar respuestas incorrectas por correctas. La segunda advierte sobre una alteración de la escala theta o de las características psicométricas de los ítems que son revisados (Revuelta et al., 2003).

Dentro del conjunto de investigaciones abocadas al estudio de la motivación del individuo en situación de examen, el impacto de la devolución inmediata de su rendimiento ha recibido una enorme atención. El *feedback* consiste en proveer al evaluado de información sobre su rendimiento (e.g. cantidad de respuestas correctas) a medida que va respondiendo el examen. Las investigaciones llevadas a cabo con tests convencionales no encontraron resultados concluyentes respecto de la relación entre el feedback y el rendimiento. En una reseña efectuada por Vispoel (1998), el autor encontró estudios que reportaban tanto efectos positivos, como negativos o nulos del feedback en el rendimiento. Los especialistas en motivación justifican estos hallazgos asegurando que la reacción del evaluado depende de múltiples variables de la situación y de la personalidad de quien responde.

En el terreno de los TAIs, el problema surge cuando se intenta determinar qué información debe brindarse al individuo mientras responde. El nivel de habilidad estimado no es un indicador adecuado porque la mayoría de los evaluados no está al tanto de las particularidades de las escala theta con la que se mide el constructo. Tampoco puede resultar del todo conveniente presentar el número de fallos cometidos. La particularidad que presenta el algoritmo de selección de los ítems del TAIs es que, independientemente del nivel de habilidad, todos los individuos tienen una similar proporción de aciertos (alrededor del 50%). Para quien está acostumbrado a la concepción tradicional que vincula la mayor cantidad de preguntas correctamente respondidas con un mayor nivel de habilidad, informar la tasa de fallos puede repercutir negativamente en su motivación. Ponsoda, Olea, Rodríguez y Revuelta (1999) suponen que esta particularidad de la administración adaptativa produce una menor sensación subjetiva de éxito que no beneficia a los individuos con mayor nivel de habilidad. En este sentido, Revuelta et al. (2003) encontraron que los evaluados tendieron a percibirse con un rendimiento peor al que habían obtenido realmente.

Los ensayos realizados para incrementar la sensación éxito percibido apuntan a manipular la dificultad de los TAIs a partir de permitir tasas de aciertos superiores a las que se utilizan corrientemente. Como sugirieron Lunz y Bergstrom (1994) posiblemente se pierda un margen de precisión en la medida pero es a costa de ganar una mayor aceptación del procedimiento por parte de los evaluados.

CONSIDERACIONES FINALES

Los TAIs han trascendido el ámbito de la investigación y han alcanzado el contexto de aplicación pero aún se deben resolver serios problemas teóricos y prácticos. La mayoría de las investigaciones referenciadas en este trabajo forman parte de la experiencia de diferentes especialistas que utilizaron TAIs para evaluar rendimiento o habilidades en un contexto determinado (laboral y académico). Por esto mismo, los resultados de estos estudios están cercenados, en alguna medida, en su posibilidad de generalización. Es posible que en otros ámbitos de aplicación donde se evalúen diferentes variables existan variantes en las reacciones de los examinados u otras conductas que no fueron todavía contempladas.

Los intentos actuales por mejorar las condiciones de aplicación se centran principalmente en cuidar la motivación de quien responde y evitar un nivel de ansiedad que obstaculice la tarea. El objetivo de los especialistas es diseñar pruebas adaptativas que, conservando en lo posible los beneficios de tipo psicométrico, no muestren desventajas motivacionales y afectivas adicionales. Esto es, lograr mayor similitud con los tests convencionales en las condiciones de administración sin perder la eficiencia en la precisión de las puntuaciones. Los progresos en la evaluación no pueden implicar perdidas de los beneficios ya adquiridos por los evaluados en las pruebas convencionales.

En definitiva, el peligro de los TAIs es que por asegurar una mayor confiabilidad de los puntajes, se atente contra la validez de la prueba. De nada sirve medir con precisión una habilidad cuando el nivel estimado está siendo distorsionado por la incidencia de ansiedad o la desmotivación del evaluado. Como aseguró Muñiz (2005), la validación rigurosa de un TAI supone un gran desafío. Será una tarea del constructor del TAI demostrar que ni la informatización ni el algoritmo adaptativo impiden que el evaluado tenga un rendimiento óptimo. Y será una responsabilidad del psicólogo evaluador exigir los estudios de validez pertinentes antes de utilizar el TAI.

BIBLIOGRAFIA

CHANG, H. & YING, Z. (1996). A global information approach to Computerized Adaptive Testing. Applied Psychological Measurement 20, 213-229.

GARRISON, W. M. & BAUMGARTEN, B. S. (1986). An Application of Computer Adaptive Testing with Communication Handicapped Examinees, Educational and Psychological Measurement 46; 23 - 35.

HAMBLETON, R.K. (2004). Theory, methods and practices in testing for the 21st century. Psicothema, 16, 696-701.

HUFF, K. L. & SIRECI, S. G. (2001). Validity issues in computer-based testing. Educational Measurement: Issues and Practice, 20, 16-25.

LUNZ, M. E., BERGSTROM, B. A. & WRIGHT, B. D. (1992). The effect of review

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