V Congreso Internacional de Investigación y Práctica Profesional en Psicología XX Jornadas de Investigación Noveno Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires, 2013.

Selección de ítems para ser utilizados como base de un Test Adaptativo Informatizado.

Lozzia, Gabriela y Abal, Facundo Juan Pablo.

Cita:

Lozzia, Gabriela y Abal, Facundo Juan Pablo (2013). *Selección de ítems para ser utilizados como base de un Test Adaptativo Informatizado. V Congreso Internacional de Investigación y Práctica Profesional en Psicología XX Jornadas de Investigación Noveno Encuentro de Investigadores en Psicología del MERCOSUR. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires.*

Dirección estable: https://www.aacademica.org/000-054/903

ARK: https://n2t.net/ark:/13683/edbf/bMq

Acta Académica es un proyecto académico sin fines de lucro enmarcado en la iniciativa de acceso abierto*. Acta Académica fue creado para facilitar a investigadores de todo el mundo el compartir su producción académica. Para crear un perfil gratuitamente o acceder a otros trabajos visite: https://www.aacademica.org*.

SELECCIÓN DE ÍTEMS PARA SER UTILIZADOS COMO BASE DE UN TEST ADAPTATIVO INFORMATIZADO

Lozzia, Gabriela; Abal, Facundo Juan Pablo
UBACyT, Facultad de Psicología, Universidad de Buenos Aires - Agencia Nacional de Promoción Científica y Tecnológica

Resumen
Se presentan los criterios utilizados para seleccionar los ítems que conformaron el Banco de Ítems de Analogías Verbales (BI-AV) de modo que sirviera de base para el desarrollo de un Test Adaptativo Informatizado (TAI). Se disponía de un conjunto de 96 reactivos unidimensionales calibrados en 3 etapas a partir del Modelo Logístico de Tres Parámetros de la Teoría de Respuesta al Ítem. Gracias al método de anclaje de ítems las estimaciones de los parámetros de estos ítems estuvieron en la misma escala aunque fueron obtenidas de 3 muestras diferentes de 547, 552 y 849 estudiantes universitarios. La selección de los ítems para ser incorporados al BI-AV se efectuó según los siguientes criterios: a) unidimensionalidad, b) ausencia de funcionamiento diferencial por género, c) buen ajuste del modelo, d) un nivel de acierto por azar cercano al esperable y e) capacidad discriminativa adecuada. El BI-AV contó sólo con los ítems más informativos ya que estos son los utilizados en los TAIs. El BI-AV quedó conformado por 64 ítems que permiten evaluar con precisión los niveles de habilidad comprendidos entre -1.75 y 3.00. Resultó más informativo para los niveles medios y altos de la habilidad para reconocer analogías verbales.

Palabras clave
Banco de Ítems, Test Adaptativo Informatizado, Teoría de respuesta al ítem, Modelo Logístico de tres parámetros

Abstract
ITEMS SELECTION TO BE USED AS THE BASIS FOR A COMPUTERIZED ADAPTIVE TEST
This work presents the criteria used for the item selection of the Item Bank of Verbal Analogies (IB-VA), to serve as a basis for the development of a Computerized Adaptive Test (CAT). It contained a set of 96 unidimensional items calibrated throughout 3 stages with the Three Parameter Logistic Model of Item Response Theory. Through the Anchor Item Method the item parameters estimates were in the same scale even though they were obtained from 3 different samples of 547, 552, and 849 college students, respectively. The item selection for their incorporation into IB-VA was performed according to the following criteria: a) unidimensionality, b) Absence of gender-related differential item functioning, c) good model fit, d) a guessing parameter level close to the expected value, and e) suitable discrimination power. The IB-VA only contained the most informative items as these are the ones used for CATs. IB-VA consists of 64 items that assesses accurately when skill levels are between -1.75 and 3.00. It was more informative for central and high levels of the ability to recognize verbal analogies.

Key words
Item Bank, Computerized Adaptive Tests, Item response theory, Three-parameter Logistic Model

Los Tests Adaptativos Informatizados (TAIs) son pruebas para la evaluación psicológica o educativa, cuyos ítems se seleccionan mediante un algoritmo computacional a partir del nivel de rasgo que progresivamente va manifestando la persona al responderlo. Es decir, su característica distintiva es que la evaluación se va adaptando al examinado (Wainer et al., 2000). Esto permite una medición más precisa presentando el menor número posible de ítems (Olea, Abad y Ponsoda, 2002). Ésta es su mayor ventaja. La misma ha sido demostrada en variadas investigaciones que indican que, a pesar de ser en promedio un 50% más corto que un Test Convencional (TC), posee igual o mayor nivel de precisión (Embretson y Reise, 2000). A su vez, redunda en los beneficios de ahorro del tiempo invertido (reduce los problemas de fatiga, desatención, aburrimiento, apatía y descuido) y en la satisfacción de los evaluados, ya que al enfrentarse a pruebas acordes con su nivel se minimizan los aspectos frustrantes que lleva aparejada toda evaluación (Abal, Lozzia, Aguerri y Galibert, 2010).

El hecho de que solo se administren los reactivos más informativos para la persona rompe el formato de los TCs (todos los evaluados responden el mismo test). A pesar de que el conjunto de ítems administrado sea diferente en cada oportunidad, se obtendrá como resultado el nivel de rasgo que le corresponde al evaluado. Esto se logra gracias a los procedimientos de la Teoría de Respuesta al Ítem (TRI) que permiten evaluar a las personas en un determinado rasgo sin necesidad de utilizar los mismos reactivos para todas ellas y expresar en la misma métrica las puntuaciones obtenidas mediante distintos conjuntos de ítems, lo cual era imposible en el marco de la Teoría Clásica de Tests (TCT) (Muñiz, 1997).

El funcionamiento de los TAIs se basa sobre dos componentes: un Banco de Ítems (BI) calibrados a partir de uno de los modelos de la TRI y un algoritmo adaptativo informatizado que ejecuta los procedimientos de inicio, estimación del nivel del rasgo, selección de reactivos y finalización que permiten la implementación del TAI. La concreción de un TAI requiere en primer lugar la creación del BI a partir del cual el TAI seleccionará los ítems a presentar en cada administración según la capacidad que va manifestando el evaluado (Wainer y Mislevy, 2000). No se trata sólo de disponer de un conjunto de ítems sino que es imprescindible conocer las características psicométricas de cada uno de los reactivos para realizar la selección adecuada. Sin embargo, no cualquier propiedad psicométrica de los reactivos es útil en el contexto de los TAIs. La posibilidad de comparar los resultados obtenidos de la administración de distintos conjuntos de ítems solo se alcanza cuando los parámetros de los reactivos se estiman mediante alguno de los modelos de la TRI. Además, a partir de los parámetros de la TRI se puede obtener la precisión con la que cada reactivo contribuye en la estimación del rasgo (la Función de Información del ítem, FI). Esta información se ha convertido en el principal criterio de selección de los reactivos por presentar en un TAI. En consecuencia, un TAIs requiere de un

Banco con ítems calibrados desde un Modelo de la TRI, es decir, de un conjunto de reactivos que miden una misma variable, que puede ser un rasgo o dominio de conocimiento, y cuyos parámetros deben estar estimados en una misma escala (calibrados) mediante un modelo de la TRI determinado (Barbero, 1996). Además, algunas de las características más importantes del TAI estarán condicionadas por el BI (e.g., el rango de valores del nivel de rasgo que permite evaluar adecuadamente y la precisión alcanzada en la estimación de los distintos niveles del rasgo, la necesidad de balance de contenido, el criterio de finalización). Por ello se dice que de la calidad del BI dependerá la calidad del TAI.

El BI es una base de datos donde se almacena cada ítem (enunciado, opción correcta, opciones incorrectas), sus características psicométricas (parámetros estimados a partir de uno de los modelos de la TRI e índices de la TCT) y la información que se pueda considerar relevante (e.g., veces que el ítem ha sido administrado, cómo ha sido creado, componentes del rasgo que mide, distribución de respuestas en los distractores). Como señalan Wainer y Mislevy (2000), el BI debe incluir ítems con adecuado poder discriminativo que evalúen en todos los niveles del rasgo. Una vez que el BI para medir un determinado constructo esté disponible, será posible desarrollar a partir de éste tantas pruebas a medida como sean necesarias.

El tamaño de un BI (cantidad de reactivos) estará sobre todo determinado por cuestiones prácticas como la finalidad y la longitud de los tests por construir, el número de personas por evaluar, la distribución del rasgo en la población por examinar, la cantidad de subdimensiones que tenga el BI (para cada una debe haber ítems que recorran todos los niveles de dificultad), las restricciones que se establezcan en el armado de los tests (e.g., tasa de exposición máxima, balance de contenidos u otros aspectos de los ítems), el peligro de divulgación de los reactivos, entre otros (Renom, 1993).

Sin embargo, lo que importa en un BI no es el tamaño en términos generales sino la distribución de sus ítems en cuanto al parámetro de dificultad. De nada sirve tener cientos de ítems que no sean apropiados a los objetivos de la evaluación ya que habrá ítems muy usados y otros que no son administrados nunca. En la mayoría de los casos, lo más conveniente es que la distribución de los parámetros de dificultad de los ítems sea similar a la del rasgo en la población de examinados (Bergstrom y Lunz, 1999). Es decir, debe haber más ítems apropiados para los niveles de θ más frecuentes en los evaluados. Por ejemplo, cuando la distribución del rasgo es normal los TAIs utilizan más ítems de dificultad intermedia mientras que los de dificultades extremas son generalmente infrautilizados.

Asimismo, la cantidad de reactivos dependerá, también, de la calidad de los mismos. Cuanto mayor sea la información que proporciona cada ítem, menos se necesitarán para alcanzar la precisión objetivo del test. Por tanto, se pueden hacer tests más cortos, dejando ítems disponibles para otras aplicaciones. Si se dispone de reactivos de muy buena calidad se puede decidir entre construir tests con la mayor precisión posible o con una precisión aceptable que requeriría administrar menos ítems y bajaría la tasa de exposición de los mismos (Xing y Hambleton, 2004). Por otro lado, si la distribución de los reactivos en cuanto al parámetro de discriminación es heterogénea (i.e., hay ítems de discriminación alta, moderada y baja) es muy probable que sólo se utilicen los de más elevada discriminación.

Teniendo en cuenta estas cuestiones, el objetivo de este trabajo es presentar los criterios que sirvieron para seleccionar los ítems que conformarían el Banco de Ítems de Analogías Verbales (BI-AV) de modo que sirviera de base para el desarrollo de un Test Adaptativo Informatizado. Se contaba con un conjunto de 96 ítems de analogías verbales calibrados en 3 etapas a partir de la TRI (para más detalles ver Lozzia, 2012). A continuación se describen brevemente los análisis efectuados en la calibración de estos ítems y los resultados obtenidos. Luego se especifican los criterios utilizados para seleccionar los reactivos adecuados para ser parte de un TAI y se presentan las características psicométricas del BI-AV así obtenido.

Método

Participantes

Se utilizaron tres muestras de 547, 552 y 849 estudiantes del segundo año de la Facultad de Psicología de la Universidad de Buenos Aires (en total 1940 individuos). Las muestras presentaron una distribución en cuanto a la edad y al género muy similar. El porcentaje de mujeres estuvo entre 84 y 87 y la mediana en edad osciló entre 20 y 21 años.

Materiales y Procedimiento de Administración

Cada una de las Pruebas de Analogías Verbales consistió en un test convencional con el formato de lápiz y papel que contenía entre 30 y 38 ítems de analogías verbales. El procedimiento de anclaje de ítems requirió que el instrumento estuviera compuesto por nuevos reactivos y por otros que tuvieran sus parámetros ya calibrados de manera que todas las estimaciones de los parámetros de los ítems estuvieran en la misma escala. Como datos adicionales se solicitaba el género y la edad de los participantes. Los participantes respondieron de forma anónima y voluntaria el instrumento en grupos reducidos sin tiempo límite.

Análisis de datos

Se llevó a cabo el análisis clásico de las características psicométricas de cada una de las tres pruebas. El Análisis estadístico en el marco de la TRI incluyó:

· *Estudio de la Unidimensionalidad* por medio del diagrama de autovalores (Scree Plot) de la matriz de correlaciones tetracóricas que proporciona MicroFact™ (Waller, 1995) y el test de Stout que proporciona DIMTEST™ (Stout, Nandakumar, Junker, Chang y Steidinger, 1991).

· *Estudio del Funcionamiento Diferencial de los Ítems (DIF) entre géneros*. Se realizó mediante el test normal para la diferencia de los parámetros de dificultad de la TRI (*b*) entre las poblaciones en cuestión. El programa BILOG-MG™ (Zimowski, Muraki, Mislevy y Bock, 1996) brinda la diferencia de los *b*s estimados con su error estándar para cada ítem.

· *Calibración de los Ítems con el Modelo de Tres Parámetros* (ML3P) por el método de Máxima Verosimilitud Marginal con el programa XCALIBRE™ (Assessment Systems Corporation, 1997). Se evaluó la bondad de ajuste del modelo sobre la base de los residuales, según su valor absoluto fuera menor a 2 (Muñiz, 1997). Se calculó la FI de cada ítem.

Selección de los ítems que formarían el Banco de Analogías Verbales

Los TAIs requieren ítems de alta calidad, por ello el BI en el que se basa debe estar compuesto por los ítems más informativos. Por ello, la selección se efectuó según los siguientes criterios: a) unidimensionalidad, b) ausencia de DIF, c) buen ajuste del modelo, d) un nivel de acierto por azar cercano al esperable y e) capacidad discriminativa adecuada (i.e., parámetro $a > 0.65$, ver Baker, 2001, pp.33-34). Finalmente se obtuvieron las características psicométricas del BI-AV (parámetros, función de información, etc.).

Resultados

La calibración de los ítems en el marco de la TRI se llevó adelante con todos los reactivos exceptuando tres ítems señalados como

con posible DIF y tres que presentaban inadecuados índices clásicos. Para todos los ítems calibrados se verificó el ajuste del ML3P. De los 96 reactivos evaluados se seleccionaron en total 64 ítems calibrados unidimensionales, que no evidenciaron DIF por género, presentaron un buen ajuste del modelo, un nivel de acierto por azar cercano al esperable y una capacidad discriminativa adecuada (parámetro $a > 0.65$).

En la Tabla 1 se muestran las características psicométricas del BI-AV. Al eliminar del BI los ítems menos discriminativos, mejoró la distribución del parámetro *c* y de los residuos. Esto se debió a que los ítems con baja discriminación estaban asociados a valores más altos tanto en el parámetro *c* como en los residuos. Si bien los residuos de estos ítems eran menores a dos, presentaban un mayor desajuste comparado al resto de los reactivos analizados. Por lo tanto, los reactivos que componen el BI-AV poseen en término medio un buen poder de discriminación (todos los ítems correspondieron a la categoría de discriminación moderada según Baker, 2001), un nivel de dificultad medio a medio-alto, un nivel de acierto por azar cercano a lo esperable para ítems con cuatro opciones de respuesta y un adecuado ajuste del ML3P.

Tabla 1. *Propiedades psicométricas del Banco de Ítems de Analogías Verbales*

	Índices del Modelo de Tres Parámetros			
	a	*b*	*c*	Residuo
Media	0.83	0.20	0.24	0.59
Desvío	0.14	0.98	0.02	0.23
Mínimo	0.65	-2.42	0.20	0.12
Máximo	1.16	2.35	0.26	1.08

Nota. a = Parámetro de Discriminación; *b* = Parámetro de Dificultad; *c* = Parámetro de Aciertos por Azar.

La distribución de los parámetros de dificultad *b* de los ítems que componen el BI-AV se ajustó a una distribución normal de acuerdo con la prueba de Kolmogorov-Smirnov ($Z = 0.66$; $p = .77$). Existe un número alto de ítems (más de ocho) para los diferentes intervalos de dificultad entre -1 y 1.50. El 61% de los ítems tuvieron valores *b* por encima de cero, lo que indicaría que el BI-AV cuenta con una mayor cantidad de ítems con niveles medios a altos de dificultad y escasos ítems en el extremo bajo de dificultad (sólo hay dos reactivos con $b < -1.50$).

En cuanto a las medidas locales de precisión, la FI del BI-AV alcanzó un valor máximo 13.74 en torno a $\theta = 0.60$. Como es de esperar dada su relación inversa, el error de medida alcanzó en este punto su valor mínimo de 0.27. En términos del coeficiente clásico de confiabilidad correspondería a 0.93. La Tabla 2 exhibe los valores que alcanza la FI del BI-AV para algunos niveles de la habilidad para reconocer analogías verbales.

Tabla 2. *Funciones de Información del Banco de Ítems de Analogías Verbales.*

	Nivel de Rasgo θ										
	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3
FI	2.36	4.52	7.44	10.35	12.58	13.70	13.26	11.30	8.47	5.64	3.39
EEE	0.65	0.47	0.37	0.31	0.28	0.27	0.27	0.30	0.34	0.42	0.54
confiabilidad clásica*	0.58	0.78	0.87	0.90	0.92	0.93	0.92	0.91	0.88	0.82	0.71

Nota. EEE: error estándar de estimación de θ. * aproximación obtenida mediante 1 - EEE^2

La FI mostró que el BI permitiría obtener estimaciones del rasgo con precisión adecuada en el intervalo de los niveles θ de -1.75 a 3.00. Para este rango de valores la FI del BI-AV es mayor a 3.30, equivalente a una confiabilidad clásica de 0.70. Mientras que el error en la medición crece hacia los niveles más bajos de θ.

Discusión

El BI-AV alcanzó las características necesarias para ser utilizado como base de un TAI. Para ello, se realizaron los análisis pertinentes para asegurar la alta calidad de los ítems del BI. La importancia de examinar el ajuste del ML3P marca una diferencia sustancial entre la TCT y la TRI, ya que por medio de éste se puede comprobar el cumplimiento de sus supuestos. Cuando el modelo se ajusta a los datos es posible suponer que la CCI representa de forma apropiada la relación entre el rasgo y la probabilidad de responder correctamente al ítem. Por otra parte, el análisis del DIF fue de suma importancia ya que las palabras que se utilizan en los ítems podrían hacer referencia a temáticas vinculadas al género y, de esta manera, favorecer a un grupo sobre otro. En efecto, el BI-AV quedó compuesto por ítems unidimensionales calibrados en la misma escala de medida que evidenciaron no tener DIF en función del género.

Aunque desde la TCT se requiere que los ítems discriminen en torno al valor central del rasgo medido debido a que utiliza indicadores globales, los TAIs necesitan que los reactivos del BI en que se sustenta discriminen en todo el espectro del constructo incluidos los extremos. Esto es lo que permite una evaluación que se adapte al nivel de habilidad que manifiesta el evaluado. Por este motivo, los criterios de selección de los reactivos en el marco de la TCT (que lleva a descartar aquellos cuya contribución no sea elevada respecto del valor medio) fueron seguidos con prudencia para no eliminar ítems con dificultades extremas. La calibración a partir del ML3P permitió conocer la contribución de cada ítem en la medición de los distintos niveles de la habilidad para reconocer analogías verbales. Esto muestra la importancia de utilizar las medidas locales de precisión que brinda la TRI.

El BI-AV contiene una cantidad suficiente y variada de ítems que permite evaluar con precisión los niveles de habilidad comprendidos entre -1.75 y 3.00. La FI del BI indicó el nivel de precisión que se podría alcanzar al evaluar cada uno de los niveles de rasgo posibles. Si bien esta función no se mantuvo uniformemente alta para todos los valores, sólo presentaba inconvenientes en el extremo inferior de la habilidad ($\theta < -1.75$). Esto se debió a que la distribución de los parámetros de dificultad no fue uniforme y hubo menos ítems que aportaran máxima información en los niveles más bajos de la variable. El BI-AV resultó más informativo para los niveles centrales y altos que para los niveles muy bajos de la habilidad. Sin embargo, esto no sería un problema ya que es muy poco probable encontrar evaluados con niveles tan bajos de habilidad en la población meta.

La obtención de un BI unidimensional, calibrado a partir de la TRI (ML3P), con una FI adecuada y libre de DIF permitió desarrollar el TAI de Analogías Verbales. Conseguir un BI con estas características fue imprescindible para el desarrollo del TAI. Por un lado, sólo los reactivos calibrados a partir de la TRI permiten las mediciones invariantes que sustentan a los TAIs. Por el otro, tanto las violaciones al supuesto de unidimensionalidad como la presencia de ítems con DIF podrían afectar la validez y equidad del TAI, ya que se estarían evaluando otros constructos además del pretendido (Segall y Moreno, 1999). La relevancia de la FI del BI reside en que la calidad del TAI depende de la calidad de los ítems que conforman el BI. Esto significa que el TAI nunca aportará más precisión para cada nivel del rasgo de

la que obtiene el BI en el que se basa. Asimismo, la incidencia de ítems defectuosos es mayor en los TAIs que en los TCs no sólo porque administran menos reactivos (Hambleton, Zaal, et al., 1991) sino también debido a que si se detecta un ítem problemático no puede ser eliminado a posteriori ya que su presencia en el TAI determinó el siguiente ítem a presentar (Potenza y Stocking, 1997).

BIBLIOGRAFIA

Abal, F., Lozzia, G., Aguerri, M. y Galibert, M. (2010) La Evaluación mediante Tests Adaptativos Informatizados. Experiencia Subjetiva del Examinado. Memorias del II Congreso Internacional de Investigación y Práctica Profesional en Psicología, Facultad de Psicología, Universidad de Buenos Aires, 4, 429-431.

Assessment Systems Corporation (1997) XCALIBRE Marginal Maximum-Likelihood Estimation [Computer program]. St. Paul, Minesota: Autor.

Baker, F. (2001) The Basics of Item Response Theory. Wisconsin: ERIC.

Barbero, M. (1996) Banco de ítems. En J. Muñiz (Ed.), Psicometría (pp. 139-170) Madrid: Universitas.

Bergstrom, B. y Lunz, M. (1999) CAT for certification and licensure. En F. Drasgow y J. B. Olson-Buchanan (Eds.), Innovations in computerized assessment (pp. 67-91) Mahwah, NJ: LEA.

Embretson, S.E. y Reise, S. (2000) Item Response Theory for Psychologists. Mahwah, NJ: LEA.

Hambleton, R., Zaal, J. y Pieters, J. (1991) Computerized adaptive testing: Theory, applications and standards. En R. Hambleton y J. Zaal (Eds.), Advances in educational and psychological testing: Theory and applications. Boston: Kluwer Academic Publishers.

Lozzia, G. (2012) Construcción de un Banco de Ítems de Analogías Verbales y su aplicación a la elaboración de un Test Adaptativo Informatizado. Tesis Doctoral no publicada. Universidad de Buenos Aires, Argentina.

Muñiz, J. (1997) Introducción a la Teoría de Respuesta a los Ítems. Madrid: Pirámide.

Olea, J., Abad, F. y Ponsoda, V. (2002) Elaboración de un banco de ítems, predicción de la dificultad y diseño de anclaje. Metodología de las Ciencias del Comportamiento, Vol. Especial, 427-430.

Potenza, M. y Stocking, M. (1997) Flawed items in computerized adaptive testing. Journal of Educational Measurement, 34, 79-96.

Renom, J. (1993) Tests Adaptativos computerizados: Fundamentos y aplicaciones. Barcelona: PPU.

Segall, D. y Moreno, K. (1999) Development of the Computerized Adaptive Testing Version of the Armed Service Vocational Aptitude Battery. En F. Drasgow y J. Olson-Buchanan (Eds.), Innovations in computerized assessment (pp. 35-65).Mahwah, NJ: LEA.

Stout, W., Nandakumar, R., Junker, B., Chang, H. y Steidinger, D. (1991) DIMTEST [Computer Program]. Champain, IL: Department of Statistics, University of Illinois.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. y Thissen, D. (2000) Computerized Adaptive Testing: A Primer (2ª ed.) Mahwah, NJ: LEA.

Wainer, H. y Mislevy, R. (2000) Item response theory, calibration, and proficiency estimation. En H. Wainer (Ed.), Computerized Adaptive Testing: A Primer (pp 61-100) Mahwah, NJ: LEA.

Waller, N. (1995) MicroFACTTM: A Microcomputer Factor Analysis Program for Dichotomous and Ordered Polytomous Data and Mainframe Sized Problems [Computer program]. St. Paul: Assesment Systems Corporation.

Xing, D. y Hambleton, R. (2004) Impact of Test Design, Item Quality, and Item Bank Size on the Psychometric Properties of Computer-Based Credentialing Examinations. Educational and Psychological Measurement, 64, 5-21.

Zimowski, M., Muraki, E., Mislevy, R. y Bock, R. (1996) BILOG-MGTM: Multiple-group IRT analysis and test maintenance for binary items [Computer program]. Chicago, IL: Scientific Software International.