CONSTRUCCION DE UN BANCO DE ITEM DE RAZONAMIENTO VERBAL *

Gabriela Susana **Lozzia** **, María Silvia **Galibert** ***, María Ester **Aguerri** **** y Horacio Félix **Attorresi** *****

Resumen

El objetivo de este trabajo es presentar el desarrollo de un *banco de ítem* de razonamiento verbal a partir de la *Teoría de Respuesta al Item* (TRI). Se presenta la TRI y su aplicación en la elaboración de bancos de ítem que posibilitan el diseño de *tests adaptativos*. Los ítem son de elección múltiple y miden la habilidad para reconocer y discriminar relaciones entre palabras. Un *banco de ítem* es un conjunto de ítem que miden

* Esta investigación fue financiada con subsidios de la Universidad de Buenos Aires (UBACyT P054 y P605), del Consejo Nacional de Investigaciones Científicas y Técnicas (PIP Nº 2426) y PICT Nº 4704 de la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT).

** Licenciada y Profesora en Psicología. Ayudante Regular de la Cátedra de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA). Becaria en el Proyecto PICT Nº 4704 de la Agencia Nacional de Promoción Científica y Tecnológica. Lope de Vega 1507, Dpto. 2 (1407) Buenos Aires, República Argentina. E-Mail: glozzia@psi.uba.ar

*** Magister Scientiae en Biometría. Profesora Adjunta Regular de la Cátedra de Matemática y Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA). Codirectora del proyecto UBACyT P054 e investigadora en los proyectos: PIP Nº 2426 del CONICET y PICT Nº 4704 de la ANPCyT.
E-Mail: galibert@psi.uba.ar

**** Magister Scientiae en Biometría. Profesora Adjunta Regular de la Cátedra de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA). Codirectora del proyecto UBACyT P605 e investigadora en los proyectos: PIP Nº 2426 del CONICET, PICT Nº 4704 de la ANPCyT y P054 de UBACyT.
E-Mail: maguerri@psi.uba.ar

***** Licenciado en Ciencias Matemáticas. Profesor Titular de la Cátedra de Estadística de la Facultad de Psicología de la Universidad de Buenos Aires (UBA). Director de los proyectos: PIP Nº 2426 del CONICET, PICT Nº 4704 de la ANPCyT y P605 de UBACyT e investigador en el proyecto UBACyT P054. E-Mail: hattorre@psi.uba.ar

una misma variable y cuyos parámetros están calibrados (estimados) en una misma escala. La construcción de un banco es un proceso de creación-calibración de ítem que se realiza en sucesivas etapas. Como los sujetos de las muestras son diferentes en cada etapa, los ítem a calibrar deben ser administrados junto con un pequeño grupo de ítem calibrados en etapas anteriores, los cuales sirven de enlace para que todas las estimaciones resulten en la misma escala. La estimación de los parámetros se lleva a cabo por el método de máxima verosimilitud marginal ajustando el *modelo logístico de tres parámetros* con el programa XCALIBRE™. Los análisis del funcionamiento diferencial (*Differential Item Functioning* - DIF) se basan en el test normal para la diferencia de los parámetros de dificultad, dicha diferencia con sus errores estándar para cada ítem es proporcionada por BILOG-MG™. Se eliminan aquellos que no ajustan al modelo y los que presentan DIF. El banco cuenta hasta el momento con 93 ítem.

Palabras clave: Banco de ítem - tests adaptativos - Funcionamiento Diferencial del Item - modelo logístico de tres parámetros - Teoría de Respuesta al Item - ítem de relaciones.

Abstract

One of the advantages of the *Item Response Theory* (IRT) compared to the Classical Test Theory is the possibility of building measurement instruments with properties independent to the subjects to be measured with them. Within IRT it is verified that the difficulty and discrimination item parameters remain invariant no matter which subject population is used. An *item bank* is a set of items which measure the same variable and whose parameters are calibrated, that is estimated, in a common scale. Whenever an item bank is available, administrations of *adaptive tests* become possible; this means the bank offers the possibility of choosing those items which assess each subject more accurately. Also, it allows designs of tests with pre-established characteristics according to the measurement objective. The application of the IRT is still incipient in

Argentina and *Differential Item Functioning* (DIF) studies have not been known so far. The aim of this work is to show the development of a Bank of Verbal Reasoning Items according to the IRT. This theory, as well as its application to the elaboration of item banks, are here in presented. These items measure the ability to identify and discriminate relationships among words. Such ability is related to the ideative factor of the verbal comprehension, which is common to all tasks of deductive, serial and probabilistic reasoning, classification and problem solution. Therefore, it is one of the main factors to get the intellectual aptitude profile of students. Each item consists of a pair of basic words with some relationship existing between them and four alternative pairs of words. The instruction commands to choose the pair with the most similar relationship to the one existing between the words in the basic pair. The construction of this bank has involved an iterative process of invention-calibration of the items. In the invention stage, the items have been selected and improved through successive pilot trials. In the selection processes, some classical statistics procedures were considered such as the question evaluation index. The calibration stage has been developed in two main phases: the initial phase and the current one. In the initial phase the calibration was done with samples from different student populations. Although some of them had responded different item sets, the calibration was done in a simultaneous running of the program BILOG-MG™ in order to obtain the estimates in a common scale. The assumption of unidimensionality was checked through a scree plot implemented by MicroFACT™. The current phase consists of adding new items to the bank until it has a substantial number of them. Before calibration, DIF analysis have been done in both phases. Taking into account the different subjects of the corresponding samples, it is necessary to manage the items together with a small group of already calibrated ones. These items serve as a link so that the parameter estimates of the new items result in the same scale as the one of the rest of the bank. Link items have been chosen with difficulty parameters representing different levels of ability and high discrimination parameters. DIF analysis have been based on the normal test of the difference between the difficulty parameters, which has been ob-

tained from BILOG-MG™. This software as well as XCALIBRE™ have been used to calibrate the items by fitting the *three-parameter logistic model*. Stout's Test of essential unidimensionality analysis has been done by using DIMTEST. After removing those items which present DIF or where the model does not fit, 93 items have been remained for the bank.

Key words: Item bank - adaptive tests - Differential Item Functioning - three parameter logistic model - Item Response Theory - relationship items.

Introducción

Características generales de los modelos de la Teoría de Respuesta al Item (TRI)

La Teoría de Respuesta al Item (TRI) constituye un nuevo enfoque para la medición psicológica y educativa que ha dado lugar a un significativo avance en la tecnología para la construcción y análisis de los tests. Entre ellos pueden mencionarse las funciones de información de los ítem y del test, errores típicos de medida distintos para cada nivel de la variable medida o el establecimiento de bancos de ítem con parámetros estrictamente definidos. Estos últimos posibilitan la construcción de tests adaptados al nivel del examinado, permitiendo exploraciones exhaustivas y rigurosas en función de las características de los sujetos.

Los modelos de la TRI formulan la existencia de una relación funcional entre los valores de la variable que miden los ítem y la probabilidad de dar una respuesta determinada, función cuya gráfica se denomina *Curva Característica del Item* (CCI). La variable por medir suele ser un rasgo que no es directamente observable, como por ejemplo algún nivel de habilidad o actitud; por lo que se la ha denominado *rasgo latente*. En los modelos más simples de la TRI este rasgo latente se considera unidimensional, es decir, se representa como una variable que toma valores en la recta real, los cuales determinan totalmente la probabilidad de elegir cada una de las posibles respuestas para el ítem. Por ejemplo, para un ítem que mide algún

tipo de habilidad, la probabilidad de respuesta correcta de dos sujetos será la misma si y sólo si dichos sujetos son igualmente hábiles.

En la TRI cada ítem queda caracterizado por sus propios parámetros, independientemente de cómo se distribuya el rasgo latente en la población en la que es administrado y de cuáles sean los parámetros de los demás ítem. Asimismo la medida del sujeto no depende de los parámetros de los ítem ni de la medida de los demás sujetos. Esta independencia entre el instrumento de medición y los sujetos es la diferencia esencial entre los enfoques de la TRI y de la Teoría Clásica de Tests (TCT).

Uno de los supuestos de los modelos de la TRI es la *unidimensionalidad de los ítem*, por la cual todos los ítem empleados en un proceso de medición o que forman parte de un banco miden el mismo rasgo, es decir, la respuesta a un ítem es explicable por una sola aptitud o actitud. Sin embargo, como es difícil que este supuesto se satisfaga totalmente ya que múltiples factores pueden afectar a las respuestas a un test, sólo se exige que haya un rasgo fundamental, como factor dominante, que explique las respuestas de los sujetos y las relaciones entre los ítem. Los modelos que requieren este supuesto se denominan *unidimensionales*. Cabe señalar que también existen *modelos multidimensionales*, cuando se hace referencia a más de un factor para explicar el rendimiento en el test.

Otro supuesto es la *independencia local*. Significa que dado un nivel del rasgo que se desea medir, digamos de habilidad, las respuestas a distintos ítem son estadísticamente independientes; es decir, que la probabilidad de responder correctamente a un ítem no aumenta ni disminuye conociendo la respuesta de un sujeto a otro ítem y su nivel de habilidad.

Modelo logístico de tres parámetros

Uno de los modelos de la TRI para describir las respuestas binarias (o dicotómicas) a los ítem para la medición de habilidades es el logístico de tres parámetros, los cuales representan tres características: el *nivel de dificultad del ítem*, su *potencia discriminatoria* y la *probabilidad de contestarlo bien por azar*. Su formulación es la siguiente:

$$P_i(\grave{e}) = c_i + \frac{1 - c_i}{1 + exp(-1.7a_i(\grave{e} - b_i))} \quad \grave{e}\mathring{a}R$$

donde:

θ es el rasgo latente que se desea medir con el ítem *i*. El hecho de que toma valores en la recta de los números reales indica que el rasgo es considerado *unidimensional*. El origen y unidad de su escala son arbitrarios, es decir, la escala está indeterminada. Suelen tomarse como cero y uno respectivamente.

P_i (θ) es la probabilidad de contestar correctamente el ítem *i* para un nivel dado de θ. El gráfico de esta función se llama *Curva Característica del Item* (CCI) y claramente, es creciente (ver Figura 1).

b_i es el *índice de dificultad del ítem i.* Coincide con el valor θ necesario para tener probabilidad $.5 + c_i/2$ de contestar correctamente el ítem *i*.

a_i es el *índice de discriminación del ítem* y es proporcional a la pendiente de la recta tangente en el punto de inflexión de la curva característica del ítem, salvo el factor $1-c_i$.

c_i es el *índice por azar del ítem i.* Es el valor de la asíntota cuando θ tiende a $-\infty$.

Cuando $c_i = 0$ se tiene el modelo de dos parámetros. En este caso el índice de dificultad b_i corresponde al nivel de θ necesario para tener una probabilidad de .5 de contestar correctamente el ítem y a_i resulta proporcional (con proporción 1.7) a la pendiente de la recta tangente en el punto de inflexión.

Si además todos los índices de discriminación coinciden, pueden considerarse iguales a 1, con lo que resulta el modelo de un parámetro o modelo de Rasch.

Un ítem será más difícil que otro si se requiere un mayor nivel de habilidad para tener la misma probabilidad de responderlo correctamente; de allí que *b* exprese el índice de dificultad del ítem, ya que puede interpretarse como el nivel de habilidad requerido para tener .5 de probabilidad de respuesta correcta en los modelos de uno o dos parámetros y $.5 + c/2$ en el modelo de tres parámetros.

Cuanto mayor sea *a*, más variará la probabilidad de respuesta correcta por unidad de cambio en el nivel de capacidad θ, lo que le da sentido a su interpretación como índice de discriminación. No obstante, esta capaci-

dad discriminatoria se da para los valores de θ que están en torno al índice de dificultad, lo cual tendrá importantes consecuencias en la construcción de tests, pues según la zona de θ que sea de interés discriminar, se elegirán unos ítem u otros.

Finalmente, c refleja la probabilidad de contestar correctamente por azar; por cuanto a niveles nulos de habilidad ($-\infty$) corresponde alguna probabilidad $c > 0$ de respuesta correcta.

Si se considera un conjunto de n ítem, el índice i variará de 1 a n: $i = 1,..., n$ y el supuesto de unidimensionalidad de los ítem quedará expresado en el modelo en que θ es el mismo para todo i. En cuanto a la independencia local, significa que la probabilidad de una determinada respuesta al ítem i para un valor dado de θ coincide con la probabilidad de dicha respuesta al ítem i para el mismo valor de θ y de respuesta a cualquier subconjunto de ítem $i_1, i_2, ..., i_k$.

En lenguaje matemático: si $U_i = u_i$, $i = 1, 2, ... n$ es la variable Bernoullí asociada a la respuesta al ítem i, entonces:

$$P(U_i = u_i / \grave{e}) = P(U_i = u_i / \grave{e}, u_{i_1}, ... u_{i_k})(i \neq i_1, ..., i_k)$$

o equivalentemente,

$$P(U_{i_1} = u_{i_1}, ..., U_{i_k} = u_{i_k} / \grave{e}) = P(U_{i_1} = u_{i_1} / \grave{e}) ... P(U_{i_k} = u_{i_k} / \grave{e})(i_j \neq i_k \text{ si } j \neq k)$$

Por tanto, la CCI queda determinada cuando se especifican estos tres parámetros, los cuales se estiman por métodos de máxima verosimilitud. El proceso de estimación de los parámetros de un ítem se llama *calibración*.

Además de la curva característica, cada ítem proporciona una *función de información* sobre el rasgo, que indica para qué niveles del mismo el ítem proporciona mediciones más precisas. Tiene que ver, por tanto, con el concepto de confiabilidad de la TCT; sólo que en la TRI un ítem no será más o menos confiable en términos absolutos sino para determinados niveles de la escala. Si el rasgo que se está midiendo es una habilidad, la función de información de un ítem fácil será mayor para niveles de habilidad bajos y menor para niveles altos; o sea que tal ítem será más apto para medir a individuos menos hábiles que para individuos más hábiles. La función

de información de un test se define como la suma de las funciones de información de los ítem que lo componen.

Ventajas del enfoque de la TRI con respecto al de la TCT

De la independencia entre los parámetros de cada ítem y de los sujetos se siguen las siguientes ventajas de la TRI con respecto a la TCT:

a.- Fijada una escala para el rasgo latente, se obtiene la misma estimación (salvo fluctuaciones del muestreo) del nivel de un sujeto cuando es medido por diferentes subconjuntos de ítem aun cuando éstos difieran en sus índices de dificultad.

b.- La estimación del nivel de un sujeto depende sólo de sus respuestas a los ítem y no de la media y el desvío estándar de la misma en la población de sujetos.

c.- Los parámetros de los ítem son invariantes entre diversas poblaciones de sujetos. En efecto, si la respuesta a un ítem sólo depende del nivel del rasgo que se desea medir, sujetos igualmente hábiles tendrán la misma probabilidad de respuesta correcta independientemente de la población de pertenencia. De este modo ya no será necesario apelar a un grupo normativo. Sin embargo el problema de adaptar las normas propio de la TCT se trueca, en la TRI, en el problema de verificar la invariancia de los parámetros de los ítem entre poblaciones. Si la probabilidad de contestar correctamente un ítem para un nivel dado de habilidad depende de alguna otra característica que la habilidad en cuestión, dicha probabilidad podrá variar entre las poblaciones que difieran en tal característica, con lo que el ítem resultaría *sesgado* al tener un *funcionamiento diferencial*. Claramente el funcionamiento diferencial se presenta cuando no se satisface el supuesto de unidimensionalidad. Así, uno de los problemas centrales de la TRI es el estudio del funcionamiento diferencial del ítem (*Differential Item Functioning* - DIF).

El problema de la indeterminación de la escala

Dado que la escala del rasgo latente está indeterminada y los parámetros a y b de los ítem dependen de ella, si se trabaja con distintas escalas

se obtendrán aparentemente distintos parámetros para un ítem. Sin embargo la invariancia de los mismos está garantizada porque las diversas escalas están vinculadas por una transformación afín; en otras palabras, los parámetros son invariantes *salvo transformaciones afines*. El problema práctico se presenta en la etapa de estimación cuando se utilizan muestras de diferentes sujetos, ya que especialmente si la distribución del rasgo latente difiere notoriamente entre las poblaciones de las que provienen, dichas estimaciones se obtendrán en distintas escalas y es necesario hallar la transformación afín que las vincula. Existen diversos métodos para ello, los cuales requieren que los ítem a calibrar se administren conjuntamente con un conjunto de ítem comunes en las distintas muestras (ítem de anclaje o de enlace) o bien que exista alguna submuestra de sujetos que responda a la totalidad de los ítem (anclaje de sujetos). Es importante que los ítem elegidos para el anclaje no tengan DIF. En lo posible se los elige de tal forma que sus índices de dificultad representen a los distintos niveles del rango de habilidad y tengan altos índices de discriminación.

Banco de ítem y sus aplicaciones. Los tests adaptativos

Un *banco de ítem* es un conjunto de ítem que mide una misma variable y cuyos parámetros están estimados en una misma escala. Los ítem con sus parámetros estimados son almacenados en una base de datos con el fin de integrar un sistema informatizado de evaluación. En la TCT, la construcción de dichos bancos se encontraba obstaculizada porque los parámetros de los sujetos dependían de los ítem a los cuales respondían, lo cual impedía la comparación de las puntuaciones de sujetos obtenidas a partir de distintos ítem y porque los parámetros de los ítem dependían de la muestra de sujetos usada para su estimación, lo cual impedía ponerlos en una misma escala. El desarrollo de la TRI y los avances tecnológicos que dieron lugar a los *software* requeridos para su aplicación, llevó a que se renovara el interés en el desarrollo de los bancos de ítem.

Millman y Arter (1984) y Hambleton y Swaminathan (1985) describieron interesantes aplicaciones que se siguen de la TRI cuando se dispone de un banco de ítem calibrados. A partir de él y mediante la función de información, es posible construir tests que proporcionen mediciones más precisas en determinados niveles de la escala, según los objetivos para los cuales éstos se conciben (tests con funciones de información prefijadas). Disponer de un banco permite además, administrar tests adaptativos (o a medida), esto es, seleccionar el conjunto de ítem más apropiado para la

medición de cada sujeto, es decir serán aquéllos que maximicen la función de información del test. En un test adaptativo se elige cada ítem por administrar según la respuesta del sujeto a un ítem anterior. La administración de los ítem continúa hasta que se alcanza un número de ítem previamente especificado o un valor determinado de precisión o error típico (Nunnally & Bernstein, 1994). La dificultad de cada ítem se halla en torno a la del ítem presentado con anterioridad, de modo que un sujeto al que se administra un test adaptativo nunca tendrá que responder ítem demasiado difíciles o demasiado fáciles para su nivel. De esta forma se evita la tendencia de los sujetos a contestar al azar y desmotivarse cuando los ítem superan sus conocimientos, así como el aburrimiento si los ítem son muy fáciles. Esto hace a los tests adaptativos más eficaces que los tradicionales, ya que generan medidas de mayor calidad utilizando menos ítem y de mayor precisión en todos los niveles del rasgo y no sólo en los niveles de habilidad cercanos a la dificultad promedio. Por ello, la utilización de este tipo de test implica un ahorro considerable de tiempo, disminuyendo los efectos de la fatiga y una mejora de las características psicométricas de fiabilidad y validez. La implementación práctica de los tests a medida actualmente es facilitada por programas de computación que proporcionan los llamados tests adaptativos computarizados o informatizados, acerca de los cuales han abundado los trabajos de investigación en los últimos años (López Pina, García, Sánchez Meca & Velandrino, 1990; Olea & Ponsoda, 1996; Wainer et al., 2000). Los bancos de ítem permiten flexibilizar la evaluación, ya que las puntuaciones obtenidas por los sujetos a partir de cualquier conjunto de ítem seleccionados del banco darán una medida del rasgo en la misma escala. Gracias a esto, para comparar los resultados no es necesario que todos los sujetos realicen el mismo test, sino que se puede elegir el conjunto de ítem que sea más adecuado al nivel de habilidad de cada uno ya que se dispone de información para los distintos niveles de habilidad. Asimismo, los bancos de ítem permiten utilizar eficientemente las respuestas de los sujetos debido a que cualquier conjunto de datos puede cargarse en la computadora para actualizar periódicamente las estimaciones de los parámetros de los ítem. Además, los bancos hacen posible diseñar tests de gran calidad ya que los ítem que se incluyen en el banco surgen de un proceso de depuración que elimina aquellos que, por falta de ajuste al modelo elegido, por no adecuarse su contenido al del banco o por cualquier otra razón, no son considerados pertinentes. Mediante las funciones de información disponibles a partir de un banco es posible estudiar la eficiencia relativa de dos tests; esto es, determinar cuánta más información puede aportar uno u otro acerca de los distintos niveles del rasgo

latente como también determinar la ponderación óptima de los ítem si se desea expresar el resultado de la medición en función del puntaje en un test. Estas aplicaciones son de particular interés para los profesionales de las áreas de psicología, educación y ciencias sociales. En Argentina el desarrollo y las aplicaciones de la Teoría de Respuesta al Item son aún incipientes. Algunos antecedentes de calibración de ítem pueden hallarse en Attorresi, Aguerri y Galibert (1999) y Cortada de Kohan (1998). Galibert y colaboradores (2000) muestran un ejemplo de aplicación al cálculo de pesos óptimos de los ítem. Aguerri, Zanelli, Galibert y Attorresi (2002) y Attorresi y colaboradores (2003) presentan desarrollos teóricos basados en simulación referidos al DIF. Una exposición detallada de la TRI y sus aplicaciones se encuentra en Muñiz Fernández (1997).

Objetivo

El presente trabajo tiene por objetivo mostrar la metodología empleada en la construcción de un banco de ítem originales de razonamiento verbal en el marco de la Teoría de Respuesta al Item.

Metodología

Para la construcción del banco se crearon y calibraron los ítem. Estos dos aspectos del trabajo se fueron entrelazando en sucesivas etapas ya que la construcción de un banco es un proceso iterativo de creación-calibración.

Características de los ítem

Los ítem son originales y miden la capacidad para reconocer y discriminar relaciones entre palabras. Cada ítem está formado por un par de palabras base que poseen una relación entre ellas y cuatro opciones de pares de palabras. Su resolución consiste en elegir entre las opciones, el par que presenta la relación más parecida a la que existe entre las palabras del par base.

JINETE - CABALLO	HARINA - MASA
1.- arqueólogo - museo	1.- juez - corte
2.- director - escuela	2.- agua - hielo
3.- administrador - consorcio	3.- melodía - música
4.- conductor - camión	4.- gestualidad - discurso
Clave: *d*	Clave: *c*

En las etapas de creación de los ítem se procedió del modo usual, teniendo en cuenta la opinión de jueces competentes y las propiedades psicométricas de la TCT a partir de pruebas piloto.

Modelización psicométrica

Al comenzar la construcción del banco se contaba con dos pruebas de relaciones, una de veinte y otra de cincuenta ítem, que habían sido administradas a muestras de distintas poblaciones y modelizadas con la TCT. La calibración se llevó a cabo para aprovechar toda la información disponible hasta ese momento, lo cual marcó una fase inicial en el desarrollo del banco. A partir de entonces se continuó trabajando con muestras de la misma población, lo que constituye la fase actual del desarrollo. En ambas fases se llevaron a cabo análisis del DIF antes de la calibración final.

Funcionamiento Diferencial de los Item

El estudio del DIF se efectuó a partir de las estimaciones de los parámetros proporcionadas por el programa BILOG-MG™ (Zimowski, Muraki, Mislevy & Bock, 1996). Dicho estudio consiste en un test normal para la diferencia entre las estimaciones de los parámetros de dificultad de cada ítem entre las poblaciones en cuestión. Se eligió este método en virtud de los resultados de los estudios de simulación de Aguerri (2000), los cuales muestran que la proporción de error de Tipo I cometido cuando se detecta DIF es menor que la obtenida con el método de Mantel-Haenszel que comete error de Tipo I en proporciones mayores que los establecidos por el

nivel de significación. Como sólo se estudia la significación estadística de la diferencia entre los índices de dificultad y no los de discriminación este tipo de DIF se llama *uniforme*. Se contrastan las hipótesis:

$$H_0 : \Delta b = b_F - b_R = 0$$

$$H_1 : \ddot{A}b = b_F - b_R \neq 0$$

donde b_F es el parámetro de dificultad estimado para el ítem en el *GF* y b_R lo es en el *GR*.

El estadístico de contraste *Z* se obtiene dividiendo la diferencia Δb por su error estándar. Es decir:

$$Z = \Delta \hat{b} / s_{\Delta \hat{b}} \quad \text{con} \quad s_{\Delta \hat{b}} = \sqrt{s_F^2 + s_R^2}$$

donde, bajo H_0, *Z* se distribuye aproximadamente como una normal estándar.

Fase inicial

Descripción de las muestras

La muestra que participó de este estudio estuvo integrada por alumnos de la Universidad de Buenos Aires (UBA).

La prueba de veinte ítem fue administrada a 349 egresantes del Ciclo Medio, 3.083 estudiantes del primer año posterior al Ciclo Básico Común (CBC) de la Carrera de Psicología y 761 estudiantes de carreras técnicas del CBC.

La prueba de cincuenta ítem se administró a 1.294 estudiantes de primer año posterior al CBC de la Carrera de Psicología.

Las pruebas corta y larga de relaciones tenían nueve ítem en común.

Análisis del DIF

Primeramente se analizó el DIF para los ítem de la prueba corta entre las diversas poblaciones de las que provenían las muestras: *Psicología* versus *CBC* e *Ingresantes* versus *Egresantes.* Se trabajó con un nivel de significación de .05. Se dejaron de lado nueve ítem con índices *QEX* menores de .1 a fin de favorecer el supuesto de unidimensionalidad del modelo. Por tanto el análisis del DIF se restringió a los once ítem restantes.

a.- *Psicología* versus *Carreras Técnicas del CBC:* No habiéndose detectado DIF en ningún ítem, se unieron los datos de las dos muestras en una sola correspondiente a la población de ingresantes a la Universidad.

b.- *Ingresantes a la Universidad* versus *Egresantes de la Escuela Media:* Se detectó DIF en el ítem 18 (p = .0019) con una diferencia estimada para los parámetros de dificultad de -.84 ($b_{egresantes} - b_{ingresantes}$), es decir que favorece a los egresantes pues el índice de dificultad es menor para éstos.

Habiéndose excluido el ítem 18 se unieron las dos muestras para reestimar los parámetros de los ítem insesgados y recién entonces se analizó el DIF por sexos. Resultaron con DIF el ítem 10 (p = .0355) y el ítem 15 (p = .0000), con diferencias ($\Delta b = b_{varones} - b_{mujeres}$) de -.36 y -.53, a favor de los varones.

Por otra parte, también se estudió el posible DIF por sexo para la muestra de sujetos que contestaron la prueba larga. Este estudio se restringió a los 47 ítem de correlaciones mayores que .1. Resultaron con DIF los ítem 5 (p = .0152), 12 (p = .0360) y 44 (p = .0447), con diferencias Δb entre los índices de dificultad de -.67, .88 y -.62, por lo que el ítem 12 favoreció a las mujeres, mientras que el 5 y el 44 a los varones. El ítem 5 es el 15 de la prueba corta, para el cual ya se había detectado DIF. Los ítem con DIF se presentan en el Anexo.

Calibración de los ítem

Dada la diversidad de muestras de esta fase, se utilizó el programa BILOG-MG™ porque permite calibrar simultáneamente los ítem administrados a diversos grupos en una escala común.

La calibración se restringió a los ocho ítem de la prueba corta y 42 de la larga con correlaciones *QEX* mayores que .10 y que resultaron sin DIF. La unidimensionalidad de estos últimos se exploró a partir de un diagrama de autovalores (*scree plot*) proporcionado por MicroFact™ (Waller, 1995). Para los seis ítem comunes a ambas pruebas se llevó a cabo un análisis del DIF con el fin de utilizarlos como enlace para la calibración conjunta de los 44 ítem provenientes de ambas pruebas. Al no detectarse DIF para ninguno de ellos se procedió a la calibración ajustando el modelo logístico de tres parámetros mediante BILOG-MG™. Se descartaron ocho ítem para los cuales se rechazó la bondad de ajuste a través del test χ^2 que proporciona dicho programa. BILOG-MG™ utiliza el método de máxima verosimilitud marginal para la estimación de los parámetros del ítem.

Resultados

Quedaron incorporados al banco 36 ítem cuyas funciones de información tienden a alcanzar un valor máximo hacia los niveles de habilidad medios a altos.

Fase actual

Sobre la base de los ítem calibrados en la fase inicial se procedió a incrementar el banco con la creación y calibración en etapas de nuevos ítem. En cada etapa se siguieron los siguientes pasos:

1.- Creación de nuevos ítem que fueron evaluados por jueces competentes y administrados en pruebas piloto para seleccionar los de mejores propiedades psicométricas y con índices de dificultad variados.
2.- Selección de los nuevos ítem por calibrar y los que funcionarían como enlace. Se escogieron como enlace ítem sin funcionamiento diferencial (DIF) entre sexos, de mayores índices de discriminación y con índices de dificultad que representaran todos los niveles de la escala de habilidad.
3.- A continuación se administraron a los alumnos del primer año posterior al CBC.
4.- Se verificó la unidimensionalidad mediante el Test de Stout de DIMTEST (Stout, Nandakumar, Junker, Chang & Steidinger, 1991).
5.- La calibración se llevó a cabo con el programa XCALIBRE (Assessment Systems Corporation, 1996) y los análisis del DIF entre

sexos con BILOG-MG™. Se utilizó el programa XCALIBRE porque permite estimar los parámetros de nuevos ítem manteniendo fijos los parámetros de los ítem ya ingresados al banco y funcionan como enlace. Esta es la diferencia esencial de procedimiento con respecto a la fase inicial. En ella, los ítem de enlace eran calibrados *simultáneamente* con todos los demás puesto que no había ninguna información anterior con respecto a sus parámetros. BILOG-MG™ permite hacer esto. Pero en la fase actual se desea calibrar nuevos ítem utilizando como enlace otros cuyos parámetros ya están estimados, por lo que es necesario que el programa mantenga fijas dichas estimaciones. Esta posibilidad la ofrece el programa XCALIBRE.

6.- El ajuste al modelo se analizó sobre la base de los residuales, según su valor absoluto se mantuviera inferior a 2; ya que XCALIBRE no proporciona el test χ^2 sino los residuos para cada ítem.

7.- Se eliminaron aquellos ítem que no ajustaban al modelo y los que presentaban funcionamiento diferencial. Sobre éstos, se intentó buscar posibles interpretaciones que explicaran su funcionamiento diferencial, las que serán tenidas en cuenta en la creación de los futuros ítem.

8.- Finalmente se agregaron al banco los ítem calibrados que no presentaron DIF ni falta de ajuste del modelo.

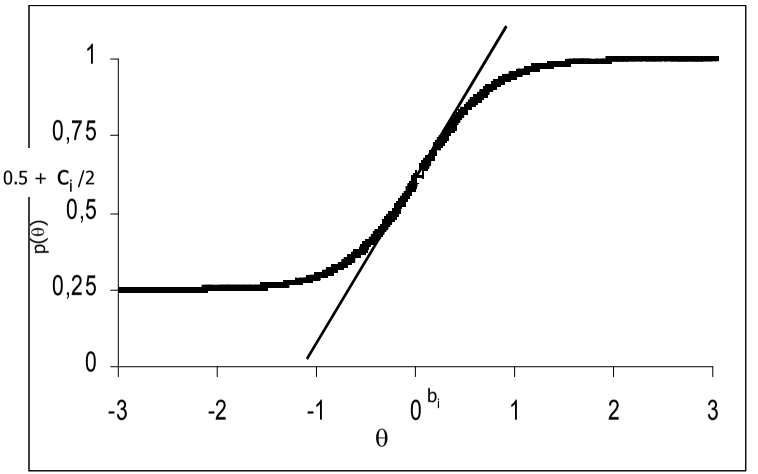
El esquema de este procedimiento se presenta en la Figura 2.

Resultados

Se detectaron dos ítem con DIF en la primera etapa. Uno de ellos, el ítem 18 que había sido calibrado como ítem 2 de la prueba larga en la fase inicial donde no había presentado DIF y estaba siendo utilizado como enlace, por lo que debió ser removido. En la segunda etapa se halló un ítem con DIF. Para ningún ítem se rechazó el ajuste del modelo, por lo tanto ingresaron al banco 19 ítem en la primera etapa, 19 ítem en la segunda y 20 ítem en la tercera y se removió uno de la fase inicial. En la Tabla 1 se exhiben los tamaños de muestra, la cantidad de ítem utilizados como enlace y por calibrar y los resultados obtenidos en cada etapa.

Hasta el momento se llevan calibrados 93 ítem originales y se proyecta continuar el proceso hasta obtener un banco de alrededor de 200 ítem.

Figura 1
Curva Característica de un Item con parámetros a, b y c



Nota:

a = 1.2
b = 0
c = .25

Figura 2
Esquema del procedimiento para la construcción de un banco de ítem

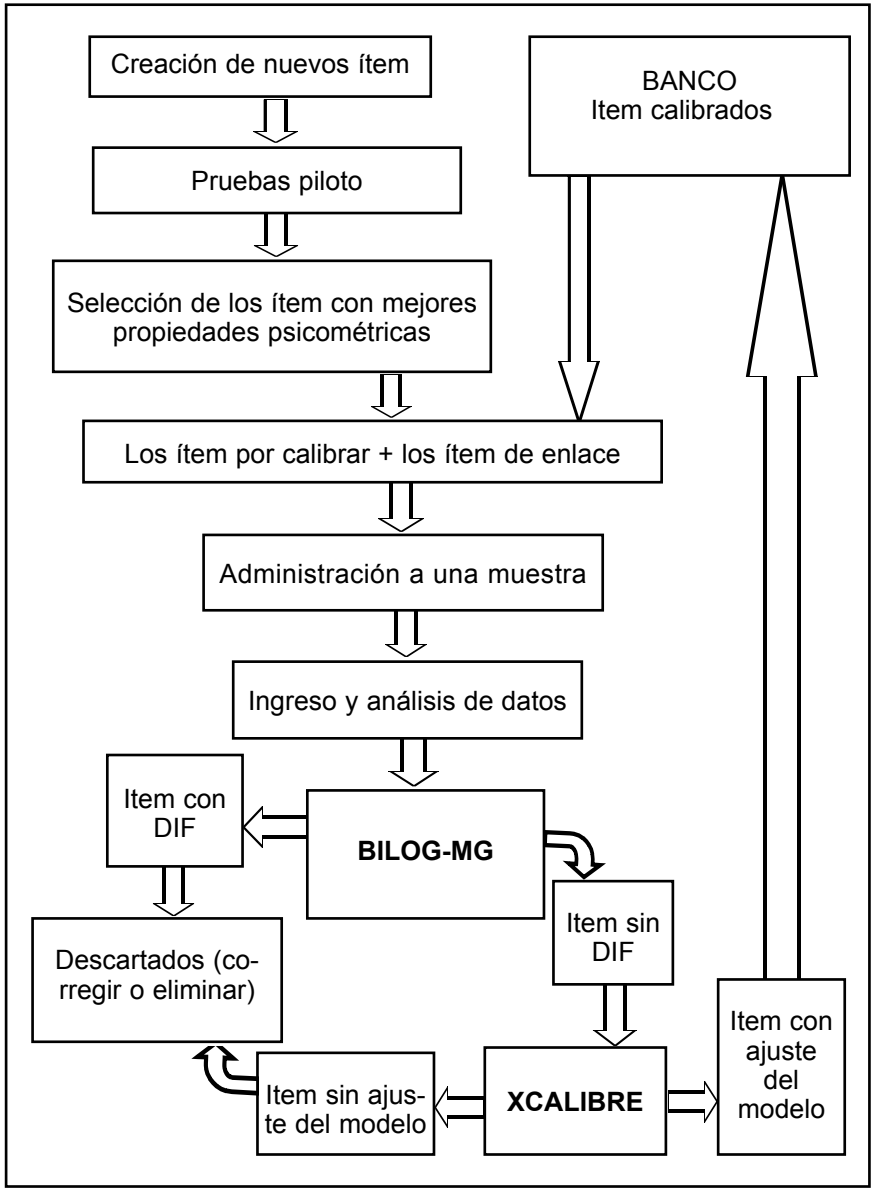


Tabla 1
Características y resultados del procedimiento de calibración en la Fase 2

Etapa	Período	Tamaño de muestra	Cantidad de ítem a calibrar	Cantidad de ítem de enlace (calibrados)	Item con DIF entre sexos, valor *p*, Δb y grupo favorecido	Item con rechazo de ajuste	Cantidad de ítem que ingresan al banco
1	1º C'01	547	20	18	Item 18 *p* = .0095 Δb = -1.14 varones Item 20 *p* = .0498 Δb = 1.08 mujeres	Ninguno	19
2	2º C'01	550	20	10	Item 20 *p* = .0400 Δb = -.52 varones	Ninguno	19
3	1º C'02	847	20	10	Ninguno	Ninguno	20

Anexo

Item con funcionamiento diferencial durante la Fase 1

Item 10 (Prueba corta)

VINO - BOTELLA
a.- agua - río
b.- leche - vaca
c.- cal - bolsa
d.- yerba - mate

Item 12 (Prueba larga)

MENTIROSO - EQUIVOCADO
a.- falso - ignorante
b.- necio - errado
c.- sincero - acertado
d.- franco - preciso

Item15 (Prueba corta)

CANOA - EMBARCACION
a.- soldado - batallón
b.- capítulo - novela
c.- daga - arma
d.- tango - canción

Item 18 (Prueba corta)

ARENA - PLAYA
a.- tierra - maceta
b.- agua - río
c.- sal - mar
d.- piedra - muro

Item 44 (Prueba larga)

JUICIO - SENTENCIA
a.- demanda - satisfacción
b.- competencia - resultado
c.- estudio - conocimiento
d.- arbitraje - reglamento

Item con funcionamiento diferencial durante la Fase 2

Item 18 (Primer cuatrimestre 2001) = Item 2 (Prueba larga)

MANTECA - LECHE
a.- cicatriz - herida
b.- italiano - latín
c.- lapicera - pluma
d.- limón - limonero

Item 20 (Primer cuatrimestre 2001)

FOTO - REVELADO
a.- bizcochuelo - horneado
b.- producto - fábrica
c.- pieza - máquina
d.- lata - reciclado

Item 20 (Segundo cuatrimestre 2001)

VIVAR - ALENTAR
a.- abuchear - desanimar
b.- burlar - desprestigiar
c.- cantar - sentimiento
d.- gritar - desalentar

Referencias bibliográficas

Aguerri, M.E. (2000). *Un estudio de simulación acerca del error de tipo I en la detección del funcionamiento diferencial del ítem* [A simulation study about the type I error for the differential item functioning]. Tesis de maestría no publicada. Universidad de Buenos Aires. Buenos Aires, Argentina.

Aguerri, M.E., Zanelli, M., Galibert, M.S. & Attorresi, H. (2002). Evaluación de un método empírico para detectar el funcionamiento diferencial del ítem [Assessment of an empirical method for the detection of the differential item functioning]. *Interdisciplinaria, 19*(2), 185-203.

Assessment Systems Corporation (1996). *User´s Manual for the Item and Test Analysis Package* [Computer program]. Saint Paul, Minnesota: Autor.

Attorresi, H., Galibert, M.S., Zanelli, M., Lozzia, G. & Aguerri, M.E. (2003). Error de tipo I en el análisis del funcionamiento diferencial del ítem basado en la diferencia de los parámetros de dificultad [Type I error in differential item functioning based on difference in the difficulty parameters]. *Psicológica. Revista de Metodología y Psicología Experimental, 24*(2), 289-306.

Attorresi, H., Aguerri, M.E. & Galibert, M.S. (1999). Aplicación del modelo logístico de tres parámetros en una prueba de completar frases [Application of the three parameters logistic model in a completion of paragraphs test]. *Investigaciones en Psicología, 4*(1), 7-25.

Cortada de Kohan, N. (1998). La Teoría de Respuesta al Item y su aplicación al Test Verbal Buenos Aires [The Item Response Theory and application in the Verbal Test Buenos Aires]. *Interdisciplinaria, 15*(1-2), 101-129.

Galibert, M.S., Aguerri, M.E. & Attorresi, H. (2000). Pesos óptimos de los ítemes en la elaboración de los puntajes [Item optimal weights in the making of scores]. *Revista Latinoamericana de Psicología, 32*(1), 79-90.

Hambleton, R. & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Boston: Kluwer-Nijholf Publishing.

López Pina, J., García, M., Sánchez Meca, J. & Velandrino, A. (1990). Test y diagnóstico psicológico por computador [Test and psychological diagnostic by computer]. En S. Algarabel & J. Sanmartín (Eds.), *Métodos informáticos aplicados a la psicología.* Madrid: Ediciones Pirámide.

Millman, J. & Arter, J.A. (1984). Issues in the item banking. *Journal of Educational Measurement, 21*, 315-330.

Muñiz Fernández, J. (1997). *Introducción a la teoría de respuesta a los ítems* [Introduction to the item response theory]. Madrid: Ediciones Pirámide.

Nunnally, J. & Bernstein, I. (1994). *Psychometric theory.* USA: McGraw-Hill.

Olea, J. & Ponsoda, V. (1996). Tests adaptativos informatizados [Computerized adaptive tests]. En J. Muñiz (Ed.), *Psicometría*. Madrid: Editorial Universitas.

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

Waller, N.G. (1995). *MicroFact™: A Microcomputer Factor Analysis Program for Dichotomous and Ordered Polytomous Data and Mainframe Sized Problems* [Computer program]. Assessment Systems Corporation.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. & Thissen, D. (2000). *Computerized adaptive testing: A primer.* (2nd. ed.). Mahwah, NJ: Lawrence Erlbaum.

Zimowski, M., Muraki, E., Mislevy, R. & Bock, R. (1996). *BILOG-MG™: Multiple-Group IRT Analysis and Test Maintenance for Binary Items* [Computer program]. Scientific Software International.

Instituto de Investigaciones
Facultad de Psicología
Universidad de Buenos Aires (UBA)
Buenos Aires - República Argentina

Fecha de recepción: 23 de mayo de 2003
Fecha de aceptación: 1 de octubre de 2003