Psicothema, 1995. Vol. 7, n.º 1, pp. 173-185
ISSN 0214 - 9915 CODEN PSOTEG

ESTIMACION DE PARAMETROS EN LA TRI: UNA EVALUACION DE BILOG EN MUESTRAS PEQUEÑAS

José A. López Pina[1]
Universidad de Murcia

En este informe presentamos un estudio de simulación para probar la exactitud de la estimación de parámetros de BILOG (V. 3.04) con muestras pequeñas (50, 100, 250 y 500) y tests cortos (10, 20, 30 y 40) en grupos donde el test fué considerado difícil ($\mu_\theta = -1$), neutro ($\mu_\theta = 0$) o fácil ($\mu_\theta = +1$). La exactitud de las estimaciones de los parámetros de los items mejoró sustancialmente con el aumento del tamaño muestral. También, la exactitud de las estimaciones de los parámetros de habilidad mejoró sustancialmente conforme se incrementó la longitud del test. Además, la estimación del parámetro de discriminación fué independiente del grado de dificultad del test. No ocurrió igual con respecto al parámetro de dificultad o de habilidad, donde BILOG fué considerablemente más exacto en sus estimaciones cuando la dificultad media del test estuvo centrada sobre la habilidad media del grupo.

Parameter estimation in IRT: an evaluation of bilog in small datasets. In this report, we present a study of simulation to prove the bias of the parameter estimation in BILOG (V. 3.04) with small groups (50, 100, 250 and 500) and short tests (10, 20, 30 and 40) in groups where the test was difficult ($\mu_\theta = -1$*), neutral ($\mu_\theta = 0$) or easy ($\mu_\theta = +1$). The bias of the item parameter estimations improve with the increase of the sample size. Also, the bias of the ability parameter estimations improve with the increase of the length test. Furthemore, the estimation of the discrimination parameter was independent of the test difficulty. It not happen the same in the difficulty or ability parameter. In these parameters, BILOG was more accurace when the test difficulty mean coincided with the group ability mean.

Introducción

La Teoría de la Respuesta al Item (TRI) es uno de los campos de investigación con mayor proyección en el ámbito de la medida psicológica y educativa. Sin embargo, a pesar de sus indudables ventajas, el proceso que emplea para estimar los parámetros de los items y de la habilidad ha dificultado una rápida implantación de la misma.

La TRI representa a un conjunto de modelos que establecen una relación no lineal entre el rasgo psicológico y la habilidad. Este tipo de relación se plasma en una función matemática, usualmente la función logística, que permite calcular la probabilidad de que un sujeto pueda acer-

Correspondencia: José A. López Pina.
Facultad de Psicología.
Campus de Espinardo. Apartado 4021
Universidad de Murcia, 30080 Murcia. Spain

(1) Parte de este informe ha sido presentado al III SIMPOSIUM DE METODOLOGIA DE LAS CIENCIAS SOCIALES Y DEL COMPORTAMIENTO celebrado en Santiago de Compostela del 12 al 16 de Julio de 1993.

tar un item conocida su habilidad. Sin embargo, la naturaleza no lineal de la función logística es uno de los obstáculos más importantes ya que dificulta enormemente el proceso de estimación de parámetros. Es decir, en la TRI, cuando queremos calibrar por primera vez un test, desconocemos tanto los parámetros de los items como los parámetros de habilidad. Por ello, es preciso recurrir a un algoritmo que nos permita obtener conjuntamente unos y otros.

Birnbaum (1968) fué uno de los primeros en proponer un método para resolver este problema. Este método se conoce como método de máxima verosimilitud conjunta y ha sido implementado en diversos programas de ordenador, tales como LOGIST (Wingersky, 1983) y ASCAL (Assessment System Corporation, 1989). El método se basa en la función de verosimilitud y su ecuación fundamental es:

$$\log(u \mid \theta, a, b, c) = \sum_{j=1}^{N} \sum_{i=1}^{n} [u_{ij} \log P(u_{ij} \mid \theta_j) + (1 - u_{ij}) \log Q(u_{ij} \mid \theta_j)] \quad (1)$$

donde u_{ij} es la respuesta dada por el sujeto j al item i. $P(u_{ij} \mid \theta_j)$ es la probabilidad de acertar el item i por el sujeto j y $Q(u_{ij} \mid \theta_j) = 1 - P(u_{ij} \mid \theta_j)$.

Aquéllos valores de los items y de la habilidad que maximizen esta función serán los verdaderos estimadores de los parámetros correspondientes. Sin embargo, resolver esta función se convierte en un enorme problema computacional ya que es preciso evaluar una ecuación de verosimilitud para cada item y habilidad. Por ejemplo, si ajustamos un modelo logístico de tres parámetros (3-p) a un grupo de tamaño N, con este método es preciso resolver simultáneamente $N + 3n - 2$ ecuaciones. Para aliviar este problema, el método de máxima verosimilitud conjunta explota la posibilidad de que los parámetros de los items y los parámetros de habilidad se puedan estimar independientemente. El proceso de estimación se realiza en ciclos. Cada ciclo se divide en dos etapas independientes. En la primera etapa se estiman los parámetros de habilidad, manteniendo fijados los parámetros de los items. En la segunda etapa se estiman los parámetros de los items, manteniendo fijados los parámetros de habilidad en sus valores actuales. Dentro de cada etapa, la estimación de parámetros opera individualmente en cada sujeto (o en cada item) y se obtienen las estimaciones a través de un algoritmo numérico que opera iterativamente conocido con el nombre de método de Newton-Raphson (Isaacson y Keller, 1966; Baker, 1992; Bunday, 1984).

El método de máxima verosimilitud no presenta problemas de convergencia cuando se conocen los parámetros de los items o los parámetros de habilidad. Sin embargo, cuando se trata de realizar una estimación conjunta de parámetros de items (parámetros estructurales) y de parámetros de habilidad (parámetros incidentales), la consistencia de las estimaciones de los parámetros estructurales se vé seriamente afectada por la presencia de los parámetros incidentales (Neyman y Scott, 1948), ya que las estimaciones de los parámetros estructurales no convergen a sus verdaderos parámetros conforme aumenta el número de parámetros incidentales (Andersen, 1980; Lord, 1980; Hambleton y Swaminathan, 1985). Para evitar este problema se han sugerido diversas soluciones. Una de ellas consiste en la utilización de procedimientos bayesianos (Swaminathan y Gifford, 1982, 1985, 1986; Hambleton y Swaminathan, 1985; Lord, 1986). Los procedimientos bayesianos permiten especificar distribuciones a priori de los parámetros que se van a estimar, con lo cual se reduce la ambigüedad de las estimaciones y se asegura la convergencia del algoritmo numérico.

Una segunda alternativa a este problema fué propuesta por Bock y Lieberman (1970). Estos investigadores separaron por completo la estimación de parámetros

de los items de la estimación de los parámetros de habilidad. Esta separación se hará efectiva a través de resolver la función de verosimilitud integrando sobre la distribución marginal de la habilidad. Una vez lograda la convergencia de los parámetros de los items se pueden obtener los parámetros de habilidad. Sin embargo, el principal inconveniente de esta aproximación reside en su carácter no cíclico. Es decir, con este método es preciso estimar todos los parámetros de una sola vez. Por ello, el proceso de estimación se restringe a tests con pocos items.

Bock y Aitkin (1981) reformularon la aproximación original utilizando una modificación del algoritmo E-M (Dempster, Laird y Rubin, 1977). Este algoritmo utiliza un conjunto de puntos cuadratura para evaluar la integral sobre la habilidad, operando, posteriormente, de forma cíclica sobre ellos. En el paso E se reemplazan los valores reales de los puntos cuadratura por valores esperados. En el paso M se resuelven las funciones de verosimilitud marginales de los parámetros de los items con los valores esperados del paso E. El conjunto E-M forma un ciclo. El proceso se interrumpe cuando no se producen cambios significativos en las estimaciones de los parámetros de los items de un ciclo a otro. Una vez que se han estimado los parámetros de los items, se estiman los parámetros de habilidad de forma independiente. En definitiva, el conjunto de parámetros que resulten de aplicar este método serán aquéllos que maximizen el logaritmo de la función de verosimilitud marginal que se define como

$$\log L_B = \sum_{m=1}^{P} C_m \log P(y_m) \qquad (2)$$

donde c_m es la frecuencia con que se presenta el patrón y_m. R es el conjunto de patrones observados y $P(y_m)$ es la fórmula de cuadratura gaussiana

$$P(y) \approx \sum_{k=1}^{q} P(y \mid Y_k)\, A(Y_k) \qquad (3)$$

donde $P(y \mid Y_k)$ es la probabilidad marginal sobre el punto de cuadratura Y_k, y $A(Y_k)$ es la ponderación positiva que corresponde a la función de densidad $g(Y)$ (Mislevy y Bock, 1990).

Este procedimiento ha sido implementado en el programa de ordenador BILOG (Mislevy y Bock, 1982, 1984, 1986, 1990). Desde su presentación, este programa, y por ende el método de estimación que propone, ha sido sometido a diversos estudios de simulación. Así, Yen (1987) realizó un estudio donde comparó BILOG vs LOGIST con el modelo logístico de 3-p. Yen empleó tests de distinta longitud (10, 20, 30 y 40 items) con un tamaño muestral fijo de 1000 sujetos y utilizó diferentes distribuciones previas para la escala de habilidad (una distribución normal y tres no normales: una distribución sesgada negativamente, otra sesgada postivamente y la tercera centrada pero platicúrtica). Yen asumió, además, distribuciones previas tanto en el parámetro de dificultad como en el parámetro de discriminación. De los resultados experimentales, Yen dedujo que BILOG casi siempre produce estimaciones más exactas en los parámetros de los items que LOGIST.

Drasgow (1989) utilizó el modelo logístico de 2-p en un estudio de simulación cuya finalidad fué comprobar cual de dos métodos de estimación (máxima verosimilitud conjunta o máxima verosimilitud marginal) produce estimaciones más exactas de los parámetros. Drasgow empleó tests y grupos de distinto tamaño. Los tests fueron de 5, 10, 15 y 25 items. Los grupos fueron de 200, 300, 500 y 1000 sujetos. Drasgow encontró que las estimaciones obtenidas con el método de máxima verosimilitud marginal fueron más

exactas que las estimaciones obtenidas con el método de máxima verosimilitud conjunta.

Seong (1990) realizó un estudio de simulación con el modelo logístico de 2-p para investigar el rol de la distribución previa de la habilidad en la estimación de parámetros. Seong empleó cuatro condiciones: el tipo de distribución previa de la habilidad (normal, positivamente sesgada, negativamente sesgada), el número de puntos cuadratura (10, 20), el tipo de distribución de θ subyacente (normal, positivamente y negativamente sesgada) y el número de sujetos (10, 1000). Seong encontró que las estimaciones de los parámetros de los items (dificultad y discriminación) fueron más exactas cuando la distribución previa de la habilidad coincidió con la distribución subyacente de θ y el tamaño muestral fué elevado. Sin embargo, el aumento de los puntos cuadratura solo mejoró sustancialmente la exactitud de la estimación cuando el tamaño muestral fué grande y ambas distribuciones de habilidad (previa y subyacente) coincidieron. Por último, la estimación de los parámetros de los items mejoró sustancialmente con el aumento del tamaño muestral.

Harwell y Janosky (1991) realizaron un estudio de simulación para investigar la exactitud con que BILOG estima los parámetros de los items bajo diversas condiciones. Estos investigadores emplearon tres condiciones: la longitud del test (15 y 25 items), el tamaño muestral (75, 100, 150, 250, 500 y 1000 sujetos) y la distribución previa del parámetro de discriminación (distribución previa ausente, $.72^2$, $.5^2$, $.25^2$, $.1^2$ en métrica lognormal). Harwell y Janosky (1991) encontraron que el efecto de la varianza previa asumida por el parámetro de discriminación mejoró sustancialmente la exactitud de las estimaciones cuando el test fué breve (15 items) y el tamaño muestral estuvo por debajo de 150 sujetos. Para tamaños muestrales mayores, el efecto de la varianza previa no mejoró significativamente las estimaciones de los parámetros de los items.

Aunque estos estudios han empleado diversas condiciones para evaluar la exactitud de los parámetros en BILOG, ninguno de ellos ha intentado evaluar esta exactitud en condiciones donde el test puede ser considerado fácil o difícil. El presente estudio de simulación, entonces, pretende evaluar el comportamiento de BILOG en condiciones extremas: tests breves y/o grupos pequeños, y tests considerados como fáciles o difíciles en un grupo determinado. El modelo logístico elegido para realizar el estudio ha sido el modelo de 2-p, ya que el modelo de 3-p requiere un elevado tamaño muestral para obtener estimaciones estables del parámetro de pseudo-azar.

Método

Condiciones experimentales

Para realizar este estudio de simulación hemos seleccionado cuatro tamaños muestrales (50, 100, 250 y 500) y cuatro longitudes de tests (10, 20, 30 y 40). Para cada uno de los cuatro tamaños muestrales se generaron tres distribuciones de habilidad normales e independientes. Una distribución tuvo $\mu_\theta = 0$ y $\theta_\theta = 1$, otra $\mu_\theta = -1$ y $\theta_\theta = 1$, y la última $\mu_\theta = +1$ y $\theta_\theta = 1$. Los parámetros de dificultad se simularon a partir de una distribución uniforme en el intervalo [–2, +2], donde ($\mu_\beta = 0$) La comparación de cada una de las distribuciones de habilidad con el correspondiente conjunto de parámetros de dificultad produjo los resultados equivalentes a un test fácil ($\mu_\theta = +1$) un test difícil ($\mu_\theta = -1$) o un test cuya dificultad se adapta perfectamente a la habilidad del grupo ($\mu_\theta = 0$). Los parámetros de discriminación se generaron a partir de una distribución uni-

forme en el intervalo [0.3, 1.6]. Todos los valores fueron generados a partir de rutinas aleatorias incluidas en SYSTAT (Wilkinson, 1990). Las distribuciones normales se generaron con la rutina ZRN y las distribuciones uniformes con la rutina URN. Los valores generados con estas rutinas se tomaron como valores verdaderos a la hora de evaluar la calidad de las estimaciones de BILOG. Estos mismos valores se utilizaron para generar las matrices simuladas de respuestas.

Simulación de matrices

Los cuatro tamaños muestrales, las cuatro longitudes del test y los tres tipos de distribución de habilidad producen una tabla de 48 celdillas. Cada una de estas celdillas corresponde a una matriz de datos resultante de combinar un tamaño muestral con la longitud del test y con un tipo de distribución de habilidad. Esta matriz se simuló a partir de un programa en TurboC siguiendo las pautas marcadas por Hambleton y Cook (1983). El programa funciona del siguiente modo. A partir de la función logística del modelo de dos parámetros, se calcula la probabilidad de acertar un item para un sujeto de habilidad θ. A continuación, esta probabilidad se compara con un número generado aleatoriamente dentro del programa. Si la probabilidad de acertar el item es mayor que el número aleatorio se considera que la respuesta del sujeto a ese item es incorrecta y, por tanto, se codifica como 0. Si, por el contrario, la probabilidad de acertar el item es menor o igual que el número aleatorio generado se considera que la respuesta es correcta y se codifica como 1. Por último, se realizaron 15 replicaciones de cada una de las celdillas, con la finalidad de obtener estimaciones estables de las medidas que permiten evaluar la exactitud de las estimaciones de BILOG. En total se analizaron 720 matrices de respuestas.

Estadísticos para evaluar el ajuste de las estimaciones obtenidas por BILOG

La exactitud de las estimaciones obtenidas con BILOG se evaluaron con tres estadísticos. El primero de ellos se conoce como error cuadrático medio (RMSE) (Hulin, Lissak y Drasgow, 1982; Skaggs y Stevenson, 1989):

$$RMSE = \sqrt{\frac{1}{201}\sum_{i=1}^{201}\left[P_i(\theta_j) - \hat{P}_j(\theta_i)\right]^2} \qquad (4)$$

Este estadístico evalúa el área entre dos CCIs generadas a partir de los parámetros verdaderos y parámetros estimados por el programa. Cuanto menor es su cuantía, más se acercan los parámetros estimados a los verdaderos.

El segundo estadístico es una medida absoluta de sesgo. Este estadístico se calcula a partir de promediar, a través de las replicaciones, las diferencias, sin tener en cuenta el signo, entre los parámetros verdaderos y los estimados.

Así, para el parámetro de discriminación, la medida absoluta de sesgo será

$$sesgo_a = \frac{\sum_{i=1}^{n} |\hat{a}_i - a_i|}{n_a} \qquad (5)$$

donde $\hat{a}_i$ es el parámetro estimado y a_i es el parámetro verdadero del item *i*. Para el parámetro de dificultad *b*, el estadístico de sesgo sin signo será:

$$sesgo_b = \frac{\sum_{i=1}^{n} |\hat{b}_i - b_i|}{n_b} \qquad (6)$$

y para el parámetro de habilidad, el estadístico de sesgo sin signo será:

$$sesgo_{\theta} = \frac{\sum_{j=1}^{n} | \hat{\theta}_j - \theta_j |}{n_{\theta}} \quad (7)$$

El tercer estadístico corresponde a una medida de sesgo equivalente a las ecuaciones 5, 6 y 7, pero teniendo en cuenta el signo. Esta nueva medida permite valorar si el programa ha sobrevalorado o infravalorado los valores verdaderos en función de que el promedio encontrado dé como resultado un valor positivo o negativo. No obstante, antes de aplicar estos estadísticos, fué necesario igualar los parámetros estimados a los parámetros verdaderos. Para ello, se utilizó el programa EQUATE descrito por Baker, Al-Karni y Al-Dosary (1991). Este programa implementa el método de curvas características de items de Stocking y Lord (1983) para obtener las constantes de igualación. Estas constantes de igualación permitirán igualar los parámetros obtenidos en BILOG a los parámetros verdaderos. Los parámetros resultantes del proceso de igualación se emplearon para obtener las medidas de RMSE y sesgo (sin signo y con signo).

Una vez obtenidas estas medidas se realizó un ANOVA para cada uno de los parámetros y medidas de sesgo con la intención de evaluar la magnitud de los efectos experimentales de las distintas condiciones empleadas. Dado que BILOG estima separadamente los parámetros de los items y los parámetros de habilidad, la estimación de los parámetros de un item no depende de la longitud del test donde esté incluido ni la estimación del parámetro de habilidad depende del tamaño del grupo donde se encuentre. Por ello, tanto en el parámetro de discriminación como en el de dificultad se realizó un ANOVA en dos sentidos de 4 (tamaños muestrales) x 3 (Distribuciones de habilidad). En el caso del parámetro de habilidad, se realizó un ANOVA en dos sentidos de 4 (longitudes de tests) x 3 (Distribuciones de habilidad). Los análisis estadísticos se realizaron con el módulo MGLH de SYSTAT (Versión 5.0).

Los resultados se exponen en una sección posterior.

Especificaciones del programa

Para realizar este estudio de simulación hemos empleado la versión 3.04 de BILOG (Mislevy y Bock, 1990). Una de las características principales de este programa es su flexibilidad. Es decir, BILOG incorpora un lenguaje de comandos que permite especificar las condiciones que se

Figura 1

Grupo de comandos empleado en un replicación con BILOG

```
FILE N50
MODELO DE 2-P
>GLOBAL DEF='C:\SIMULA\N50\DA5.DAT', NPARM=2,SAVE;
>SAVE PARM='DA5P.DAT',SCORE='DA5H.DAT';
>LENGTH NITEMS=10;
>INPUT NTOT=10,NALT-1000,INOPT=1,NTD=2;
>TEST TNAME=N50-10;
>CALIB;
>SCORE METHOD=3,NOPRINT;
```

utilizarán en la estimación de parámetros (el funcionamiento de algunos comandos ha cambiado con respecto a versiones anteriores del programa). La figura 1 presenta un ejemplo de un fichero de comandos empleado en este estudio de simulación.

Aunque no se especifica directamente en esta figura BILOG asume, por defecto, una distribución previa log-normal para la pendiente (parámetro de discriminación) con $\mu_a = 0$ y $\theta_a = .5$ y una distribución normal con $\mu_\theta = 0$ y $\sigma_\theta = 1$ en el parámetro de habilidad. BILOG, sin embargo, no asume ninguna distribución previa en el parámetro de dificultad aunque puede hacerlo si así lo precisa el investigador. Por último, para obtener las estimaciones de la habilidad se empleó un estimador modal bayesiano a posteriori (MAP).

Resultados

Recubrimiento de las CCIs

El error cuadrático medio es una medida de área que incluye todos los parámetros estimados por BILOG. Por ello, para esta medida hemos realizado un ANOVA utilizando las tres condiciones del estudio: Tamaño muestral, longitud del test y distribución de habilidad. La tabla 1 recoge los promedios de las 15 réplicas en las tres condiciones experimentales. Entre paréntesis aparece la desviación típica promedio de las 15 réplicas.

La medida de área propuesta en esta ecuación examina en qué medida la CCI obtenida a partir de las estimaciones de los parámetros de BILOG (CCI empírica) reproduce con exactitud la CCI generada a partir de los parámetros verdaderos (CCI teórica). En este sentido, cuanto menor sea esta medida, más se acercarán los parámetros estimados por BILOG a los parámetros verdaderos.

A partir de los resultados presentados en la tabla 1 y del ANOVA realizado sobre los promedios de las tres condiciones experimentales podemos afirmar que el ajuste de la CCI empírica a la CCI teórica mejoró sustancialmente con el aumento del tamaño muestral ($F_{3,669} = 992.0877, p \leq .0000$). Aunque menos apreciable, también mejoró el ajuste entre ambas curvas con el aumen-

Tabla 1
Area entre las CCIs de los parámetros verdaderos y estimados por BILOG
(Media y desviación típica)

Longitud del test	Media de la habilidad	Tamaño muestral			
		50	100	250	500
10	-1	0771(0447)	0578(0381)	0503(0320)	0406(0246)
	0	0726(0397)	0499(0267)	0358(0197)	0292(0161)
	+1	0881(0461)	0630(0443)	0519(0410)	0430(0299)
20	-1	0762(0470)	0580(0397)	0457(0376)	0297(0221)
	0	0682(0404)	0489(0300)	0341(0212)	0245(0156)
	+1	0747(0495)	0550(0357)	0396(0328)	0331(0248)
30	-1	0728(0487)	0587(0423)	0423(0367)	0306(0235)
	0	0677(0392)	0509(0311)	0367(0221)	0245(0139)
	+1	0745(0481)	0595(0440)	0418(0350)	0301(0255)
40	-1	0748(0482)	0588(0421)	0404(0291)	0300(0244)
	0	0692(0421)	0498(0306)	0341(0215)	0240(0158)
	+1	0727(0466)	0602(0414)	0399(0331)	0302(0238)

to del tamaño del test ($F_{2,669}$ = 80.5144, $p \leq$ .0000). Por último, también el grado de dificultad del test influyó decisivamente en las estimaciones de los parámetros ($F_{3,669}$ = 21.3023, $p \leq$ .0000) ya que el ajuste entre CCIs (teórica y empírica) fué mejor cuando la dificultad del test coincidió con la habilidad media del grupo que en los dos casos restantes (test fácil o difícil). Es precisamente en estas dos condiciones donde BILOG tuvo mayores problemas para obtener estimaciones exactas de los parámetros. No obstante, esta situación no se mantuvo constante a través de todas las condiciones experimentales ya que la interacción tamaño muestral y longitud del test resultó significativa ($F_{9,669}$ = 2.4854; $p \leq$ .0085). Así, de la tabla 1 se deduce que el ajuste entre las CCIs fué apreciablemente peor cuando el tamaño muestral fué de 50 sujetos y la longitud del test de 10 items que en tamaños muestrales mayores.

Sin embargo, cuando la longitud del test fué de 30 items o más y el test fácil o difícil, no se encontraron diferencias apreciables entre las CCIs, ya que los RMSE obtenidos en cada condición se solapan. Esto es un indicador de que BILOG tiene problemas a la hora de estimar parámetros en tests no centrados sobre la escala de habilidad. No obstante, no parece que haya un efecto diferencial en función del grado de dificultad del test cuando éste tiene 30 ó más items.

Sesgo en las estimaciones de los parámetros.

Parámetro de discriminación

La tabla 2 recoge los resultados correspondientes a la medidas de sesgo en el parámetro de discriminación.

Como se aprecia en esta tabla, BILOG estimó con mayor exactitud los parámetros de discriminación conforme aumentó el tamaño muestral ($F_{3,705}$ = 303.944, $p \leq$ .000). También, se encontraron diferencias ($F_{2,705}$ = 6.655, $p \leq$ .001) en las estimaciones de este parámetro en función de la distribución del rasgo latente. La interacción entre tamaño muestral y tipo de distribución de habilidad, sin embargo, no fué significativa. Parece, pués, que BILOG fué más exacto a la hora de estimar los parámetros de discriminación en distribuciones centradas sobre la dificultad del test que en distribuciones no centradas, confirmando con ello que el grado de dificultad del test afecta directamente a las estimaciones de los parámetros en la TRI. Además, estos problemas de estimación en el parámetro de discriminación se presentaron por igual en todos los tamaños muestrales e independientemente del grado de dificultad del test.

La tabla 2 recoge también los resultados correspondientes a los promedios sobre las 15 replicaciones del estadístico de sesgo con signo. Esta medida nos per-

Tabla 2
Medidas de sesgo en el parámetro de discriminación

	Distribución de habilidad					
Tamaño	Sesgo sin signo			Sesgo con signo		
muestral	-1	0	+1	-1	0	+1
50	241	234	241	008	009	005
100	195	189	191	003	017	009
250	160	138	151	009	001	015
500	128	111	125	005	015	012

mite comprobar si BILOG presentó alguna tendencia a sobrevalorar o infravalorar las estimaciones de los parámetros verdaderos. Realizado el ANOVA pertinente ninguno de los dos efectos principales (tamaño muestral y distribución de habilidad) resultaron significativos, indicando que BILOG tiende a estimar consistentemente los parámetros de discriminación ya sea con pocos o muchos sujetos y en tests fáciles, difíciles o neutros. No obstante, de los resultados de esta tabla podemos deducir cierta tendencia de BILOG a infraestimar los parámetros de discriminación.

Parámetro de dificultad

La tabla 3 recoge los resultados de los estadísticos de sesgo en el parámetro de dificultad.

Como se aprecia en la misma, la exactitud de las estimaciones del parámetro de dificultad mejoró ($F_{3,705} = 430.230, p \leq .000$) conforme aumentó el tamaño muestral. También BILOG fué más exacto a la hora de estimar los parámetros de dificultad cuando la habilidad media del grupo coincidió con la dificultad media del test que cuando no coincidieron ($F_{2,705} = 72.702$, $p \leq .000$). Sin embargo, del mismo modo que en el parámetro de discriminación, la interacción de estos dos factores no resultó significativa, indicando los problemas de estimación de BILOG se presentan por igual en un tamaño muestral bajo que alto, en un test fácil, difícil o neutro. En cuanto a la posibilidad de que BILOG sobreestimara o infraestimara los parámetros de dificultad, sólo la condición de distribución de habilidad centrada o no ha resultado estadísticamente significativa ($F_{3,705} = 22.844, p \leq .000$). Examinando los promedios de la medida de sesgo con signo de la tabla 3 podemos deducir que BILOG tendió a sobrestimar los parámetros cuando el test fué difícil mientras que tendió a infraestimarlos cuando el test fué fácil. Sin embargo, cuando el test estuvo centrado sobre la distribución de habilidad, en tamaños muestrales bajos presentó cierta tendencia a infraestimar los parámetros mientras que en tamaños muestrales elevados presentó cierta tendencia a sobreestimarlos.

En definitiva, parece que BILOG estima mejor los parámetros de dificultad cuando la dificultad media del test está centrada sobre la habilidad media del grupo. Cualquier desviación de este patrón general parece incidir en un empeoramiento de las estimaciones del parámetro de dificultad de los items. En cuanto a la dirección del sesgo no parece que el tamaño muestral haya sido una condición determinante ya que la dirección del sesgo fué prácticamente igual en todas los niveles mientras que

Tabla 3
Medidas de sesgo en el parámetro de dificultad

	Distribución de habilidad					
Tamaño	Sesgo sin signo			Sesgo con signo		
muestral	-1	0	+1	-1	0	+1
50	308	281	323	008	004	020
100	242	192	251	001	002	013
250	188	139	188	008	002	005
500	144	099	147	022	003	020

si fué una condición determinante que la distribución de habilidad estuviera centrada o no sobre la dificultad del test.

Parámetro de habilidad.

En la tabla 4 aparecen las medias y desviaciones típicas de las escalas de habilidad utilizadas en el estudio.

En términos generales, la media y desviación típica de las estimaciones de la habilidad obtenidas por BILOG se acercan bastante a la media y desviación típica de los parámetros verdaderos. Nótese, sin embargo, que las medias de las estimaciones de habilidad cuando la dificultad del test no estuvo centrada sobre la habilidad del grupo difieren en mayor cuantía que cuando test y habilidad media estuvieron centrados. Este es un claro ejemplo del efecto de regresión hacia la media que provoca la inclusión de una distribución previa informada en la función de verosimilitud marginal empleada por BILOG. Es un hecho ya ampliamente documentado (Swaminathan y Gifford, 1982; Baker, 1987) que la inclusión de una distribución previa produce el efecto de regresar las estimaciones de los parámetros en curso hacia la media de la distribución a posteriori, con lo cual es previsible que las estimaciones de los parámetros estén más concentradas alrededor de la media a posteriori de las estimaciones que si no se hubiera empleado una distribución previa para obtener las estimaciones. Nótese que las desviaciones típicas de las estimaciones, en la mayor parte de las condiciones experimentales, fueron menores que las desviaciones típicas de los parámetros verdaderos de partida.

La tabla 5 recoge los resultados de las medidas de sesgo para el parámetro de habilidad en las distintas condiciones empleadas en este estudio.

A partir de los resultados de esta tabla y del análisis estadístico pertinente se deduce que la estimación del parámetro de habilidad mejoró conforme aumentó la longitud del test ($F_{3,703}$ = 2257.851, $p \leq .0000$) (Stone, 1992). También, la exactitud ($F_{2,703}$ = 157.862, $p \leq .0000$) de las esti-

Tabla 4
Media y desviación típica de los parámetros de habilidad obtenidos por BILOG en cada condición experimental

Tamaño muestral	Media de la habilidad	Longitud del test			
		10	20	30	40
50	-1	-930(706)	-956(814)	-1.018(916)	-1.002(949)
	0	029(719)	006(874)	-013(918)	-018(957)
	+1	891(646)	918(843)	987(884)	999(932)
100	-1	-898(671)	-837(807)	-829(862)	-979(902)
	-0	051(730)	-004(863)	002(894)	-001(1.137)
	+1	895(763)	935(783)	1.000(855)	995(876)
250	-1	-633(664)	-920(779)	-845(834)	-984(883)
	0	021(713)	004(840)	000(884)	021(905)
	+1	899(578)	925(755)	981(828)	971(853)
500	-1	-839(654)	-802(766)	-962(829)	835(849)
	0	006(697)	011(812)	003(1.070)	003(888)
	+1	827(598)	927(756)	954(820)	968(854)

maciones de los parámetros de habilidad dependió de que la distribución de habilidad estuviera centrada o no sobre la dificultad del test. Además, ambos factores interactuaron significativamente ($F_{6,703}$ = 11.337, $p \leq .0000$).

Efectivamente, como es esperable, si aumenta la longitud del test también aumenta la información que disponemos sobre la habilidad latente del individuo y, por tanto, en BILOG obtenemos una mejor estimación del parámetro correspondiente. También es esperable que la estimación de la habilidad sea más exacta cuando la distribución de la habilidad está centrada sobre la dificultad del test (Seong, 1990) que cuando no lo está, ya que ese mismo resultado se obtuvo con el parámetro de dificultad y como sabemos ambos parámetros se encuentran sobre el mismo continuo. No está tan clara, sin embargo, la interacción ya que las diferencias entre los promedios en las distintas longitudes de test parecen más un efecto debido al error de muestreo por el bajo número de replicaciones que a diferencias reales entre esas longitudes.

En cuanto a la dirección de las estimaciones, no se han encontrado diferencias significativas entre las longitudes de test empleadas, aunque si entre los distintos tipos de distribución de habilidad ($F_{2,704}$ = 235.575, $p \leq .000$). También resultó significativa la interacción entre ambos efectos ($F_{6,704}$ = 235.575, $p \leq .000$) como resultado de las diferencias en sesgo tan elevadas en el segundo factor (Distribución de habilidad). Es decir, BILOG sobreestima e infraestima los parámetros tanto más severamente cuanto más breve es el test. Sin embargo, conforme aumenta su longitud, las estimaciones de los parámetros de habilidad se hacen más exactas. Este mismo patrón se repite ya sea el test fácil o difícil. Sin embargo, si la distribución de la habilidad está centrada sobre la media de dificultad del test, BILOG sólo infraestima los parámetros de habilidad cuando el test es muy breve (10 items). En los tres casos restantes, la exactitud de las estimaciones es muy elevada.

Conclusiones

En principio, la generalización de los resultados de un estudio de simulación a la realidad está muy limitada. Sin embargo, una ventaja importante de estos estudios reside en que permiten poner a prueba cualquier procedimiento o programa de ordenador en situaciones extremas que no siempre encuentran una contrapartida real. Teniendo ésto en mente, presentamos las siguientes conclusiones.

De los resultados obtenidos en este estudio se deduce que BILOG permite obtener estimaciones exactas de los parámetros de los items y de habilidad aún cuan-

Tabla 5
Medidas de sesgo en el parámetro de habilidad

	Distribución de habilidad					
Longitud	Sesgo sin signo			Sesgo con signo		
del test	-1	0	+1	-1	0	+1
10	447	413	473	-115	-025	122
20	346	297	334	-043	-003	071
30	275	260	280	-021	001	021
40	240	216	256	-018	006	017

do el tamaño muestral sea bajo y la longitud del test breve. No obstante, bajo estas dos condiciones, las estimaciones de los parámetros de discriminación y dificultad mejoran sustancialmente conforme aumenta el tamaño muestral y las estimaciones de los parámetros de habilidad también mejoran sustancialmente conforme aumenta la longitud del test.

Sin embargo, cuando se trata de estimar parámetros y utilizamos un grupo de sujetos cuya habilidad media no está centrado sobre la dificultad media del test, BILOG presenta algunos problemas. Así, las estimaciones de los parámetros de discriminación y dificultad de los items fueron mejores cuando la distribución de habilidad estuvo centrada que cuando no la estuvo. BILOG, además, tendió a infraestimar los parámetros de discriminación en todos los tamaños muestrales e independientemente de si la distribución de habilidad estuvo centrada o no, mientras que en el parámetro de dificultad, BILOG tendió a infraestimar los parámetros cuando la habilidad media fué mayor que la dificultad media del test y viceversa.

En cuanto a los parámetros de habilidad, BILOG también obtuvo las mejores estimaciones cuando la distribución de habilidad estuvo centrada sobre la dificultad media del test. Sin embargo, en las dos condiciones restantes, BILOG tuvo mayores problemas para obtener estimaciones exactas, sobreestimando e infraestimando los parámetros correspondientes. Nótese, además, que la dirección de las estimaciones de los parámetros de habilidad en cada una de estas dos condiciones fué justamente la contraria a la obtenida con respecto al parámetro de dificultad. Así, cuando la habilidad del grupo fué menor que la dificultad media del test (test difícil), BILOG tendió a sobreestimar los parámetros de habilidad e infraestimar los parámetros de dificultad mientras que en la condición opuesta ocurrió justamente lo contrario. Es decir, en un test fácil, BILOG tendió a infraestimar los parámetros de habilidad y sobreestimar los parámetros de dificultad de los items.

Referencias

Andersen, E.B. (1980). Discrete statistical models with social science applications. Amsterdam: North-Holland.

Assessment Systems Corporation (1988). ASCAL: User's manual for the 2- and 3-parameter IRT calibration program. St. Paul, MN: Author.

Baker, F.B. (1987). Methodology Review: Item parameter estimation under the one—, Two—, and Three-Parameter Logistic Models. Applied Psychological Measurement, 11, 111-141.

Baker, F.B. (1992) Item response theory: Parameter estimation techniques. New York: Marcel Dekker.

Baker, F.B., Al-Karni, A. y Al-Dosary, I. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. Applied Psychological Measurement, 15(1), 78.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F.M. Lord y M.R. Novick, Statistical theories of mental test scores (p. 397-492). Reading, MA: Addison-Wesley.

Bock, R.D. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorimh. Psychometrika, 46, 443-459.

Bock, R.D. y Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Bunday, B.D. (1984). Basic optimisation methods. London: Edward Arnold.

Dempster, A.P., Laird, N.M. y Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. Applied Psychological Measurement, 13, 77-90.

Hambleton, R.K. y Cook, L.L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. En D.J. Weiss, New horizons in testing: Latent trait test theory and computerized adaptive testing, p. 31-49. New York: Academic Press.

Hambleton, R.K. y Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Harwell, M.R. y Janosky, J.E. (1991). An empirical study of the effects of small datasets and varying prior variances of item parameter estimation in BILOG. Applied Psychological Measurement, 15(3), 279-292.

Hulin, C.L., Lissak, R.I. \& Drasgow, F. (1982). Recovery of two— and three-parameter logistic item characteristic curves: A monte carlo study. Applied Psychological Measurement, 6, 249-260.

Isaacson, E. y Keller, H.B. (1966). Analysis of numerical methods. New York: Wiley.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. Journal of Educational Measurement, 23, 157-162.

Mislevy, R.J. y Bock, R.D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items. Chicago: International Educational Services.

Mislevy, R.J. y Bock, R.D. (1984). BILOG I Maximum likelihood item analysis and test scoring: Logistic model. Mooresville, IN: Scientific Software.

Mislevy, R.J. y Bock, R.D. (1986). PC-BILOG: Item analysis and test scoring with binary logistic models [Computer Program]. Mooresville, IN: Scientific Software.

Mislevy, R.J. y Bock, R.D. (1990). PC-BILOG 3: Item analysis and test scoring with binary logistic models [Computer Program]. Mooresville, IN: Scientific Software.

Neyman, J. y Scott, E.L. (1948). Consistent estimates based on partially consistent observations. Econometrika, 16(1), 1-32.

Seong, T.J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied Psychological Measurement, 14(3), 299-311.

Skaggs, G. \& Stevenson, J. (1989). A comparison of Pseudo-Bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model. Applied Psychological Measurement, 13, 391-402. Stocking, M.L. y Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stone, C.A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. Applied Psychological Measurement, 16(1), 1-16.

Swaminathan, H. y Gifford, J.A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-192.

Swaminathan, H. y Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. Psychometrika, 50, 349-364.

Swaminathan, H. y Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. Psychometrika, 51, 589-601.

Wilkinson, L. (1990). SYSTAT: The systems for statistics (v. 5.0). Evanston IL: Author.

Wingersky, M.S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. En R.K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

AGRADECIMIENTOS: El autor agradece al editor y un revisor anónimo sus valiosos comentarios sobre la primera versión de este manuscrito.

Aceptado el 21 de junio de 1994