Revista de Psicología
Universidad de Chile

2014, 23(1), 53-70
ISSN impreso : 0716-8039
ISSN en línea : 0719-0581
www.revistapsicologia.uchile.cl

Estimación bayesiana de un modelo psicométrico multinivel con efectos aleatorios

Bayesian estimation of a multilevel psychometric model with random effects

Javier Revuelta y Carmen Ximénez

Universidad Autónoma de Madrid, Madrid, España

Resumen

El presente estudio aborda la problemática del desarrollo de modelos psicométricos para diseños de investigación multinivel, que tienen como objetivo comparar el nivel medio de los sujetos procedentes de diferentes grupos organizados en niveles definidos de forma jerárquica. Se presenta un modelo psicométrico multinivel basado en la Teoría de Respuesta al Ítem (TRI) y un procedimiento para obtener estimaciones en modelos jerárquicos de TRI mediante métodos bayesianos. El modelo se refiere a datos dicotómicos y un rasgo latente de naturaleza unidimensional, y pone el énfasis en el aspecto jerárquico del análisis. Además de presentar formalmente el modelo, se ilustra su aplicación mediante un ejemplo que incluye datos empíricos referidos a un test de conocimiento matemático aplicado a 1000 estudiantes españoles, organizados en escuelas procedentes de tres regiones. Los resultados proporcionan información sobre cada estudiante, escuela y región. Adicionalmente, se incluye el código de sintaxis empleado para la estimación bayesiana a partir de los programas OpenBUGS y Stan, con el propósito de que el lector pueda adaptar la sintaxis empleada a su propio problema. Finalmente, se discuten las implicaciones del uso de este tipo de modelos multinivel y las posibles líneas de investigación futura.

Palabras clave: Teoría de Respuesta al Ítem, modelos multinivel, estimación bayesiana.

Abstract

The present study examines the problem of the development of psychometric models for multi-level designs, that aim to compare the medium level of subjects from different groups organized in levels hierarchically defined. A psychometric multilevel model based on the Item Response Theory (IRT) and a Bayesian procedure to obtain estimations in hierarchical models of IRT are presented. The model refers to dichotomous data and a one-dimensional latent trait, and put emphasis on the hierarchical aspect of the analysis. In addition to formally introducing the model, an illustration of the application of the procedure is presented by an example that includes empirical data referred to a test of mathematical knowledge that was applied to a sample of 1,000 Spanish students organized in schools from three different regions. The results provide information about each student, school, and region. Additionally, the syntax code used in the Bayesian estimation with the OpenBUGS and Stan programs is included in order to provide the reader with a tool that can be adjusted to his/her own research problem. Finally, the implications of the use of multilevel models and future research directions are discussed.

Keywords: Item Response Theory, multilevel models, Bayesian estimation.

Este trabajo ha sido parcialmente financiado por el proyecto PSI2012-31958 del Ministerio de Economía y Competitividad de España.
Contacto: J. Revuelta. Departamento de Psicología Social y Metodología. Universidad Autónoma de Madrid, 28049, Madrid, España.
javier.revuelta@uam.es

Cómo citar este artículo:
Revuelta, J. y Ximénez, C. (2014). Estimación Bayesiana de un modelo psicométrico multinivel con efectos aleatorios. *Revista de Psicología, 23*(1), 53-70. doi: 10.5354/0719-0581.2014.32874

Introducción

El desarrollo de los modelos psicométricos y sus algoritmos de estimación ha permitido abordar diseños de investigación cada vez más sofisticados, empleando datos procedentes de tests psicológicos y escalas educativas (Martínez Arias, Hernández y Hernández, 2006). Uno de estos desarrollos consiste en los diseños multigrupo, en los que es posible comparar el nivel medio de los sujetos, por ejemplo, de diferentes países (Jöreskog, 1971). Al mismo tiempo, los diseños multigrupo plantean el problema de la invarianza, consistente en determinar si las propiedades psicométricas de los tests, por ejemplo su dificultad, permanecen constantes de un grupo a otro (Alwin y Jackson, 1981; Meredith, 1993).

Los modelos jerárquicos van un paso más allá en el desarrollo de los modelos multigrupo. En un diseño jerárquico, además de disponer de muestras procedentes de varias poblaciones, estas últimas se organizan en varios niveles cada vez más generales. Por ejemplo, los exámenes PISA (*Program for International Student Assessment*) y TIMSS (*Trends in International Mathematics and Science Study*) se aplican a estudiantes de educación primaria y secundaria, recogiéndose muestras de miles de personas de países de todo el mundo, agrupadas en escuelas y regiones. De este modo, existen varios grupos en el diseño (e.g., las escuelas, agrupadas en unidades mayores definidas por las regiones), lo cual ha conllevado la necesidad de elaborar modelos psicométricos que proporcionen información a todos estos niveles, pudiendo comparar escuelas y regiones entre sí, y obteniendo, de este modo, una información más rica y elaborada que en las aplicaciones clásicas en las que todos los sujetos proceden de una misma población (Goldstein, 2004).

El propósito de este artículo es proporcionar una guía útil de cómo puede aplicarse en la práctica un modelo psicométrico multinivel basado en la *Teoría de Respuesta al Ítem* (TRI; Fox, 2005, 2007). Nos centraremos en un modelo sencillo, en el que los datos son dicotómicos y el rasgo latente tiene naturaleza unidimensional, para poner el acento en el aspecto jerárquico. Aparte de exponer el planteamiento formal del modelo, ilustraremos su aplicación en un ejemplo empírico y mostraremos el código informático para realizar la estimación. Los datos empíricos se refieren a un test de conocimiento matemático aplicado a estudiantes españoles de educación secundaria. Los estudiantes se organizan en escuelas procedentes de tres regiones españolas. El análisis de datos se basa en un modelo integrado que proporciona información de cada estudiante, escuela y región. El propósito es que este ejemplo pueda servir de base al lector para sus propios análisis, de modo que el código informático pueda ser utilizado o modificado para adaptarlo a otros problemas similares.

El artículo se organiza del siguiente modo. En primer lugar presentamos la descripción teórica del modelo psicométrico multinivel, en el marco de la TRI, y el procedimiento para obtener estimaciones en modelos jerárquicos de TRI mediante métodos Bayesianos. A continuación, se ilustra la aplicación del modelo multinivel mediante un ejemplo que usa datos empíricos. Por último, se exponen las conclusiones y posibles extensiones del modelo propuesto. Adicionalmente, se incluyen dos apéndices en los que se explica con detalle cómo se han utilizado los programas *OpenBUGS* y *Stan*, con el propósito de que el lector pueda adaptar la sintaxis empleada a su propio problema.

Modelos multinivel

Los modelos multinivel se propusieron inicialmente en un marco estadístico general, no específico de los modelos psicométricos, y se caracterizan porque las unidades de la muestra se agrupan en varios niveles (para una introducción completa a modelos multinivel, véase Goldstein, 2003).

En el caso más sencillo solo existen dos niveles, de modo que las unidades del nivel 1 están anidadas en las del nivel 2. Un ejemplo sería el de una encuesta de intención de voto, donde las unidades de nivel 1 serían los sujetos encuestados y las de nivel 2 las regiones de procedencia; y otro sería un análisis de medidas repetidas, donde los sujetos se evalúan antes y después de una terapia; las unidades de nivel 1 serían los sujetos y las de nivel 2 el momento de recogida de datos.

Un ejemplo de modelo multinivel es la regresión lineal en varios grupos. En un análisis tradicional, la regresión lineal de Y sobre X viene dada por $y_i = a + bx_i + e_i$, donde a y b son la constante y la pendiente de la regresión, respectivamente. En un análisis multinivel los sujetos se agrupan en unidades mayores, por ejemplo, por ciudades de procedencia, y se obtiene una ecuación de regresión para cada ciudad j que viene dada por $y_{ij} = a_j + b_j x_{ij} + e_i$. De este modo, los parámetros a_j y b_j pueden variar de una ciudad a otra. En este ejemplo, las variables X e Y son variables observadas en la muestra. En un modelo psicométrico se mantiene la lógica de un muestreo con unidades anidadas, aunque la variable independiente es una variable no observada o factor latente.

La estimación de modelos psicométricos multinivel suele realizarse dentro de un marco inferencial Bayesiano, debido a las complejidades técnicas que supone su estimación desde otros enfoques. En la práctica, esto supone asumir que cada uno de los parámetros del modelo sigue una determinada distribución a priori, y realizar la estimación mediante métodos de simulación. El ejemplo que se describe en este artículo se basa en uno de los métodos más extendidos para

la estimación Bayesiana, consistente en utilizar el programa informático *OpenBUGS* (Spiegelhalter, Thomas, Best y Lunn, 2003, 2014). De modo complementario, también se proporcionan indicaciones de cómo utilizar un programa más novedoso, *Stan* (Hoffman y Gelman, 2011), que se encuentra en una fase más incipiente de desarrollo, pero que presenta buenas perspectivas.

Modelo multinivel de respuesta al ítem. Supongamos que un sujeto responde a un ítem con dos categorías de respuesta indicadas por los códigos 0 y 1 (e.g., error versus acierto, desacuerdo versus acuerdo, etc.). Según el modelo logístico de dos parámetros, la probabilidad de la categoría 1 viene dada por:

$$P_i = \frac{\exp(c_i + a_i\theta)}{1+\exp(c_i + a_i\theta)}, \quad (1)$$

donde los parámetros a_i y c_i se denominan *escala* e *intercepto*, respectivamente, y determinan la fuerza de la relación del ítem con el rasgo latente θ y la proporción de respuestas correctas cuando $\theta = 0$. En el contexto de la TRI, es usual transformar los parámetros de la ecuación (1) del modo siguiente para facilitar la interpretación del modelo (Hambleton y Swaminathan, 1985):

$$P_i = \frac{\exp(a_i(\theta - b_i))}{1+\exp(a_i(\theta - b_i))}, \quad (2)$$

donde b_i es el parámetro de dificultad y resulta de la transformación $b_i = -c_i / a_i$.

En un análisis tradicional, el modelo de TRI se estima asumiendo que todos los sujetos proceden de una única población en la cual los niveles de rasgo, θ, siguen una distribución normal (0, 1). Sin embargo, en este artículo el modelo (1) se aplica en un diseño jerárquico en el que los sujetos son estudiantes agrupados en varias escuelas, y estas últimas pertenecen a varias regiones. El análisis multinivel proporciona información a estos tres niveles: individuo, escuela y región, y permite compararlos entre sí. Los parámetros del modelo (1) o (2) corresponden al primer nivel, los individuos, que pueden compararse basándose en sus niveles de rasgo estimados. En un segundo nivel, los individuos se agrupan en escuelas, y cada una de ellas tiene un nivel de rasgo medio que permite comparar el nivel académico de los miembros de unas escuelas y otras. En concreto, el nivel de rasgo del sujeto *i*, perteneciente a la escuela *j* de la región *k* viene dado por:

$$\theta_{ijk} = \beta_{jk} + e_{ijk}, \quad (3)$$

donde β_{jk} es el nivel medio de los estudiantes de la escuela j y e_{ijk} un error aleatorio que expresa las diferencias entre los distintos estudiantes de la escuela. La distribución de e_{ijk} se asume que es normal ($0, \sigma_{jk}$), donde la desviación típica σ_{jk} indica, en parte, según veremos, cuánto varían entre sí los estudiantes de la escuela.

Como las escuelas se agrupan por regiones, en el modelo existe un tercer nivel según el cual el valor medio de cada escuela viene dado por:

$$\beta_{jk} = \gamma_k + u_{jk}, \quad (4)$$

donde γ_k es la media de la región k y u_{jk} es un error aleatorio distribuido según la normal ($0, \omega_k$), donde ω_k es el efecto de la región *k* sobre la variabilidad de los niveles de rasgo.

Finalmente, se postula un cuarto y último nivel que agrupa a todos los sujetos de todas las escuelas y regiones. Es decir, la muestra completa, y según el cual los valores medios de las regiones vienen dados por:

$$\gamma_k = \eta + o_k, \quad (5)$$

donde η es el la media de todas las regiones y o_k es un error aleatorio distribuido según la normal $(0, \delta)$.

Estadísticamente, el modelo propuesto pertenece a la clase de los modelos mixtos, que combinan efectos fijos y aleatorios (McCulloch, Searle y Neuhaus, 2008). Los parámetros θ_{ijk}, β_{jk} y γ_k constituyen efectos aleatorios y siguen una distribución normal (aunque podría haberse establecido otra). Los demás parámetros son efectos fijos porque teóricamente toman un valor único. Con respecto a los efectos aleatorios, muchas veces lo relevante es la proporción de variabilidad de los niveles de rasgo explicada por la agrupación de escuelas y regiones, que viene determinada por los parámetros σ_{jk}, ω_k y δ.

A partir de las ecuaciones (3), (4) y (5), es posible aislar los componentes que influyen en la varianza de θ_{ijk} y estimar la importancia de cada uno de ellos. En concreto, bajo el su-

puesto de que todos los errores son independientes entre sí, la varianza del nivel de rasgo es:

$$\begin{aligned} Var(\theta_{ijk}) &= Var(\beta_{jk} + e_{ijk}) \\ &= Var(\gamma_k + u_{jk} + e_{ijk}) \\ &= Var(\eta + o_k + u_{jk} + e_{ijk}) \\ &= \delta^2 + \omega_k^2 + \sigma_{jk}^2, \end{aligned} \tag{6}$$

donde los términos δ^2, ω^2_k y σ^2_{jk} son las contribuciones de cada uno de los niveles del modelo a la variabilidad de las puntuaciones. El término σ^2_{jk} representa el efecto, o contribución, de la escuela *j* de la región *k* sobre la varianza del nivel de rasgo. Esto no debe confundirse con la variabilidad dentro de la escuela. Si $\sigma^2_{jk} = 0$, la variabilidad dentro de la escuela no sería cero sino $\delta^2 + \omega^2_k$. Además, si todas las escuelas de una misma región tuviesen $\sigma^2_{jk} = 0$, todas ellas tendrían la misma varianza, por lo que no habría efecto de la escuela dentro de esa región. Del mismo modo, ω^2_k representa el efecto de la región *k* sobre la varianza. Si $\omega^2_k = 0$ entonces la varianza dentro de la región sería δ^2.

A partir de un desarrollo similar al de la ecuación (6), puede comprobarse que la covarianza entre dos niveles de rasgo es:

$$Cov(\theta_{ijk}, \theta_{i'j'k'}) = \begin{cases} \delta^2 + \omega_k^2 + \sigma_{jk}^2, & \text{si } i = i', j = j' \text{ y } k = k' \\ \delta^2 + \omega_k^2, & \text{si } i \neq i', j = j' \text{ y } k = k' \\ \delta^2, & \text{si } i \neq i', j \neq j' \text{ y } k = k' \\ 0, & \text{si } i \neq i', j \neq j' \text{ y } k \neq k' \end{cases} \tag{7}$$

Para valorar la contribución de los diferentes niveles sobre la variabilidad de las puntuaciones suelen utilizarse correlaciones intraclase o correlación entre los niveles de rasgo de dos estudiantes de la misma escuela (θ_{ijk} y $\theta_{i'jk}$), que resulta ser la proporción de varianza atribuida al efecto de la escuela:

$$\rho_1 = \frac{\delta^2 + \omega_k^2}{\delta^2 + \omega_k^2 + \sigma_{jk}^2}. \tag{8}$$

Del mismo modo, la correlación intraclase de los valores medios de dos escuelas de una misma región (β_{jk} y $\beta_{j'k}$) es:

$$\rho_2 = \frac{\delta^2}{\delta^2 + \omega_k^2}. \tag{9}$$

Estimación Bayesiana

El método más habitual para obtener estimaciones en modelos jerárquicos de TRI es la estimación bayesiana aplicada mediante simulación Monte Carlo con cadenas de Markov (Gilks, Richardson y Spiegelhalter, 1996). Este método proporciona muestras de la distribución a posteriori de los parámetros. Supongamos que **X** es la matriz de respuestas dicotómicas (e.g., acierto y error) al test de matemáticas y $\boldsymbol{\varepsilon}$ es el vector de parámetros correspondiente a los efectos fijos y que tiene como elementos a_i, c_i, η, σ_{jk}, ω_k y δ. La distribución a posteriori es proporcional a la función de densidad conjunta de todas las variables involucradas en el problema

$$\begin{aligned} f(\boldsymbol{\varepsilon} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &\propto f(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varepsilon}) \\ &= f(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon}), \end{aligned} \tag{10}$$

donde $f(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\varepsilon})$ es la función de densidad conjunta de los datos observados y de los efectos aleatorios, condicionada en el vector de efectos fijos. Supongamos que la muestra consta de *H* ítems, n_k escuelas en la región k, y n_{jk} estudiantes en la escuela j. A partir de la ecuación (1) y de las distribuciones normales involucradas en las distribuciones (3), (4) y (5), se obtiene que la densidad conjunta de los datos y efectos aleatorios es

$$\begin{aligned} &f(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\varepsilon}) = \\ &= \prod_{k=1}^{K} f(\gamma_k) \prod_{j=1}^{n_k} f(\beta_{jk}) \prod_{i=1}^{n_{jk}} f(\theta_{ijk}) \prod_{h=1}^{H} P_{ih}^{x_{ih}} (1 - P_{ih})^{1 - x_{ih}}. \end{aligned} \tag{11}$$

La distribución a priori, $f(\boldsymbol{\varepsilon})$, indica el conocimiento disponible acerca de $\boldsymbol{\varepsilon}$ antes de observar la matriz de datos. La distribución $f(\boldsymbol{\varepsilon})$ viene especificada por las funciones de distribución a priori de cada parámetro, que en este ejemplo han sido:

$$\begin{aligned} a_i &\sim \text{gamma}(0.25,\ 0.25) \\ c_i &\sim \text{normal}(0,\ 4) \\ \sigma^2_{jk} &\sim \text{gamma}(1,\ 1) \\ \omega^2_k &\sim \text{gamma}(1,\ 1) \\ \delta^2 &\sim \text{gamma}(1,\ 1) \end{aligned} \tag{12}$$

En la ecuación (12) se ha utilizado la distribución gamma como distribución a priori para aquellos parámetros que no pueden tomar valores negativos, y la distribución normal para los que sí. Los valores de los parámetros de estas distribuciones a priori son arbitrarios, en ausencia de otra

evidencia, y se han utilizado valores que proporcionen una distribución con una desviación típica elevada, como reflejo de la ausencia real de conocimiento previo acerca del valor real de los parámetros. A partir de estas definiciones, la distribución a priori de $\boldsymbol{\varepsilon}$ es:

$$f(\boldsymbol{\varepsilon}) = f(\delta^2)\left(\prod_{k=1}^{K} f(\omega_k^2)\prod_{j=1}^{J} f(\sigma_{jk}^2)\right)\prod_{h=1}^{H} f(a_i)f(c_i). \qquad (13)$$

En un problema de estimación Bayesiana, $\boldsymbol{X}$ permanece constante al valor estimado en los datos y se utiliza la función de densidad conjunta, $f(\boldsymbol{X, \theta, \beta, \gamma, \varepsilon})$, para obtener muestras de los parámetros $\boldsymbol{\theta, \beta, \gamma}$ y $\boldsymbol{\varepsilon}$. La media y desviación típica de estas muestras constituyen el estimador Bayesiano esperado a posteriori (EAP) y su error típico, calculados por simulación. El tipo de simulación que se emplea para tomar estas muestras es la simulación de Monte Carlo con Cadenas de Markov (MCMC, Gilks et al., 1996).

La simulación MCMC se caracteriza porque existe una relación estadística entre cada valor muestreado y el siguiente, que forman así una cadena de Markov. Suelen simularse varias cadenas en paralelo, cada una empieza a evolucionar desde unos valores iniciales arbitrarios hasta converger en la distribución a posteriori de la ecuación (10). Para implementar este método en la práctica es necesario utilizar cadenas de elevada longitud. La primera mitad de los valores muestreados se descarta porque se considera un periodo de arranque muy contaminado por la arbitrariedad de los valores iniciales. Para valorar si las cadenas de Markov han convergido a la distribución a posteriori, suele utilizarse el estadístico $\sqrt{R}$ de Gelman y Rubin (1992) y Brooks y Gelman (1998), que se calcula por separado para cada uno de los parámetros. La lógica de $\sqrt{R}$ es similar al concepto de proporción de varianza explicada en un diseño de ANOVA, e indica la proporción de varianza del parámetro entre las cadenas frente a la varianza dentro de las cadenas. Se considera que un valor de $\sqrt{R} \leq 1.1$ es indicador de convergencia.

Estudio empírico

Muestra

Se han analizado las respuestas a un test de matemáticas de 1000 estudiantes de bachillerato, 517 varones y 483 mujeres, con edades comprendidas entre los 16 y 17 años y de niveles socioeconómicos medios. Los datos fueron recogidos en el año 2012.

Instrumento

El test consta de 12 ítems agrupados en tres áreas de contenido: 1) funciones y ecuaciones (e.g., suma y resta de fracciones, resolución de sistemas de ecuaciones lineales); 2) cálculo y geometría (e.g., derivadas e integrales); y 3) probabilidad y estadística (e.g., cálculo de la probabilidad conjunta de varios sucesos independientes, y la probabilidad de la unión de sucesos). Los ítems constaban de cinco alternativas de respuesta, una de ellas correcta. Para el análisis se utilizaron datos dicotómicos de acierto/error.

Procedimiento de análisis y resultados

En primer lugar, se estimaron modelos con uno, dos y tres factores latentes para valorar la dimensionalidad del test. Para ello se empleó el paquete *mirt* disponible como parte del lenguaje R (Chalmers, 2012). Los resultados aparecen en la tabla 1, que contiene el logaritmo de la función de verosimilitud, el número de parámetros estimados, el estadístico chi-cuadrado de ajuste relativo entre modelos anidados G^2 y los estadísticos AIC y BIC. Al comparar los modelos con uno y dos factores, el estadístico G^2 toma un valor significativo, lo que indica que el modelo de un factor obtiene peor ajuste que el de dos. Cuando se comparan los modelos con dos y tres factores, G^2 no es significativo, por lo que se mantiene la hipótesis nula de igual bondad de ajuste para los modelos con dos y tres factores. En definitiva, la prueba chi-cuadrado de bondad de ajuste lleva a escoger el modelo de dos factores. Sin embargo, el estadístico BIC, que combina el criterio de bondad de ajuste con el de parsimonia, sugiere que el modelo de un factor sería el más apropiado entre los tres comparados porque la mejora en bondad de

Tabla 1
Bondad de ajuste de modelos con uno, dos y tres factores latentes

Factores	Log. verosimilitud	Parámetros	$G^2[df]$p-valor	AIC	BIC
1	-7469.9	24		14987.9	15105.7
2	-7456.6	35	26.8[11].01	14983.1	15154.9
3	-7451.5	45	10.1[10].43	14993.0	15213.9

ajuste que se produce al aumentar el número de factores no es lo suficientemente grande como para compensar la complejidad añadida al modelo que supone aumentar el número de dimensiones.

Con respecto al modelo jerárquico, el test se aplicó en tres provincias españolas. Puesto que el propósito del artículo es ilustrar el tipo de información proporcionada por estos modelos jerárquicos, para mantener la confidencialidad, no ofreceremos datos sobre la identidad de las regiones y, en su lugar, las denominaremos región 1, 2 y 3. Hemos tomado datos de 10 institutos de bachillerato de cada provincia, por lo que en el primer nivel tenemos los 1000 valores de θ_{ijk} de los estudiantes. El segundo nivel corresponde a las escuelas, con 10 valores de β_{jk} en cada región, y en el tercer nivel tenemos los tres valores de γ_k. En el ejemplo, $J = 10$ y $K = 3$.

Se utilizaron dos programas informáticos *OpenBUGS* y *Stan* para obtener las muestras de parámetros. En los apéndices A y B aparece el código empleado en ambos programas y unas indicaciones para que el lector pueda adaptarlos a sus propios datos. Como ambos programas proporcionaron estimadores muy similares, por simplicidad, solo se comentarán los resultados proporcionados por *OpenBUGS*.

Se han tomado muestras con cuatro cadenas de Markov en paralelo de 40000 elementos cada una, descartándose los primeros 20000 y quedando así una muestra de 80000 elementos para realizar la inferencia, al juntar la segunda mitad de las cuatro cadenas. El valor de $\sqrt{R}$ fue menor de 1.1 para cada uno de los parámetros al cabo de las 40000 muestras, por lo que de acuerdo con este criterio, la convergencia es adecuada.

La figura 1 muestra la evolución de ocho parámetros (a_1, c_1, β_{11}, σ^2_{11}, γ_{11}, ω_{11}, η_{11} y δ_{11}) a lo largo de las 20000 muestras correspondientes a la primera cadena tras eliminar el periodo de arranque. En las gráficas se aprecia que no existe una evolución con tendencia ascendente o descendente para ninguno de los parámetros, sino que oscilan de forma aleatoria en torno a un valor promedio que permanece constante a medida que avanza el número de muestras. La ausencia de una tendencia ascendente o descendente en estas gráficas es una indicación de que la cadena ha convergido a la distribución posterior.

Los programas informáticos proporcionan 80000 valores muestreados para cada parámetro, al juntar las cuatro cadenas. Con estos valores se han elaborado los histogramas que aparecen en la figura 2, y que corresponden a los mismos ocho parámetros de la figura 1. Estos histogramas son las distribuciones a posteriori de los parámetros obtenidas por simulación.

La tabla 2 contiene los estimadores puntuales EAP para el modelo de TRI de las ecuaciones (1) y (2) junto con sus errores típicos. Los parámetros del último ítem se han fijado a unos valores constantes arbitrarios para fijar la escala de la variable latente. Existen otros métodos para fijar la métrica, como estandarizar los valores de rasgo muestreados en cada iteración del procedimiento de muestreo (Fox, 2010). Sin embargo, después de varias pruebas prácticas, se ha encontrado que lo más conveniente es utilizar este método para fijar la métrica y no imponer restricciones en θ.

En la tabla 2 puede advertirse que los parámetros de los ítems no están diferenciados por escuelas y regiones. A la hora de estimar el modelo, se ha asumido invarianza de los parámetros para poder obtener estimaciones comparables de los niveles de rasgo, junto con sus medias y varianzas, en las distintas escuelas y regiones. El modelo multinivel también permite análisis más flexibles en los que se permite que los parámetros de los ítems varíen de una escuela, o de una región, a otra. Esto proporcionaría una base estadística para investigar las causas sustantivas que subyacen a las diferencias de rendimiento entre grupos, en lo que se conoce como análisis del funcionamiento diferencial de los ítems o del test (Osterlind y Everson, 2009).

Tabla 2

Parámetros estimados del modelo de medida. Estimador esperado a posteriori y error típico a posteriori

Ítem	*a***	Se(a)	c**	Se(c)	b
1	0.69	0.11	0.43	0.09	-0.62
2	1.11	0.15	-0.08	0.12	0.07
3	0.73	0.11	-0.63	0.10	0.86
4	0.82	0.11	-0.54	0.11	0.66
5	1.42	0.18	-1.20	0.15	0.85
6	1.11	0.14	-0.82	0.13	0.74
7	0.28	0.08	-0.99	0.09	3.54
8	0.44	0.09	-0.50	0.09	1.14
9	0.65	0.10	-0.21	0.10	0.32
10	0.79	0.11	-1.66	0.13	2.10
11	0.29	0.09	0.87	0.09	-3
12*	1		0		0

Nota:

* Los parámetros del ítem 12 se han fijado a valores constantes para establecer la métrica del rasgo latente.

** *a* y *c* son los parámetros de escala e intercepto, respectivamente, y *b* es el parámetro de dificultad del ítem, calculado como $b = -c/a$.

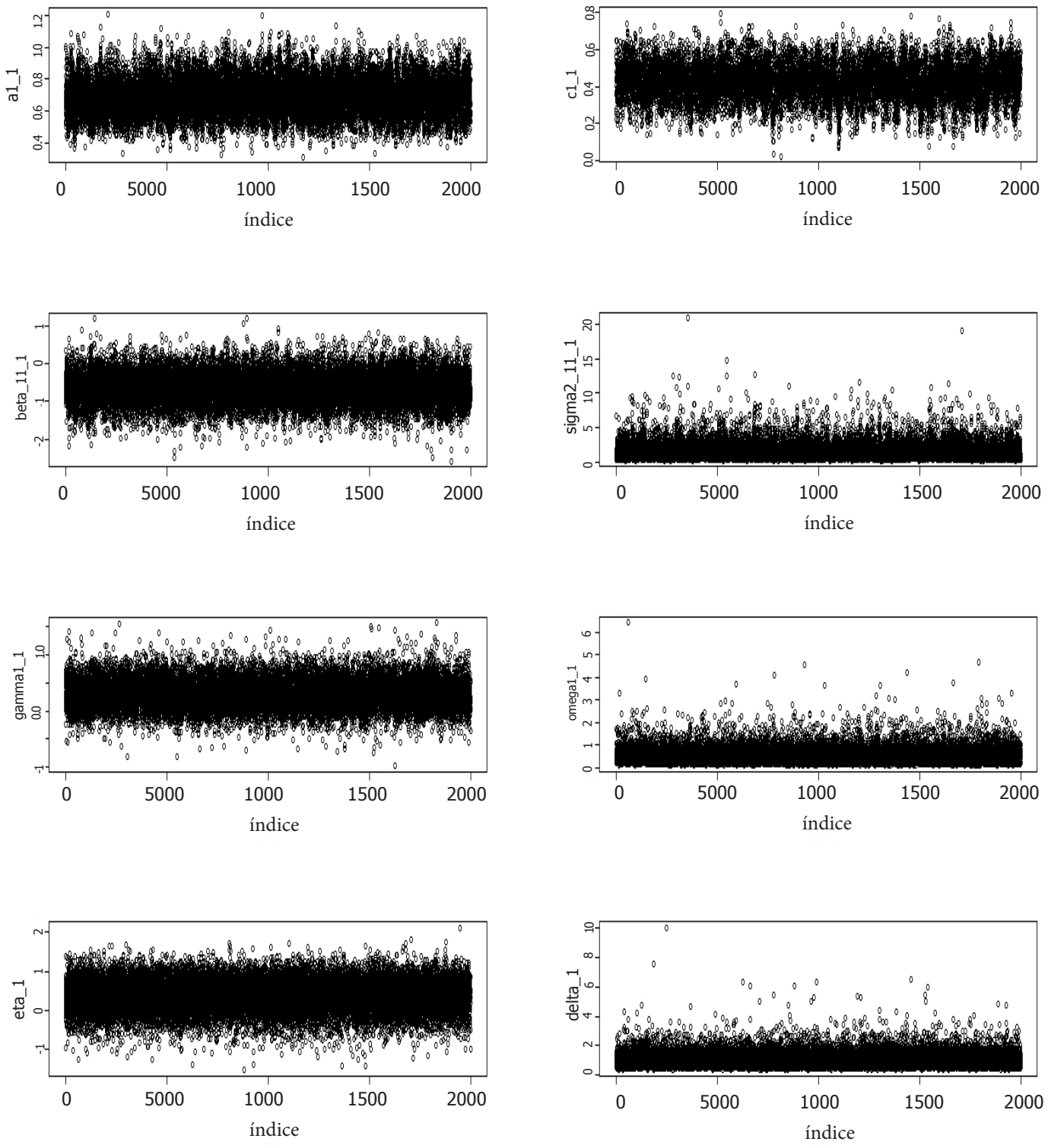

Figura 1. Evolución de algunos elementos del vector de parámetros $\boldsymbol{\varepsilon}$ a lo largo de las 20000 muestras de la primera cadena de Markov.

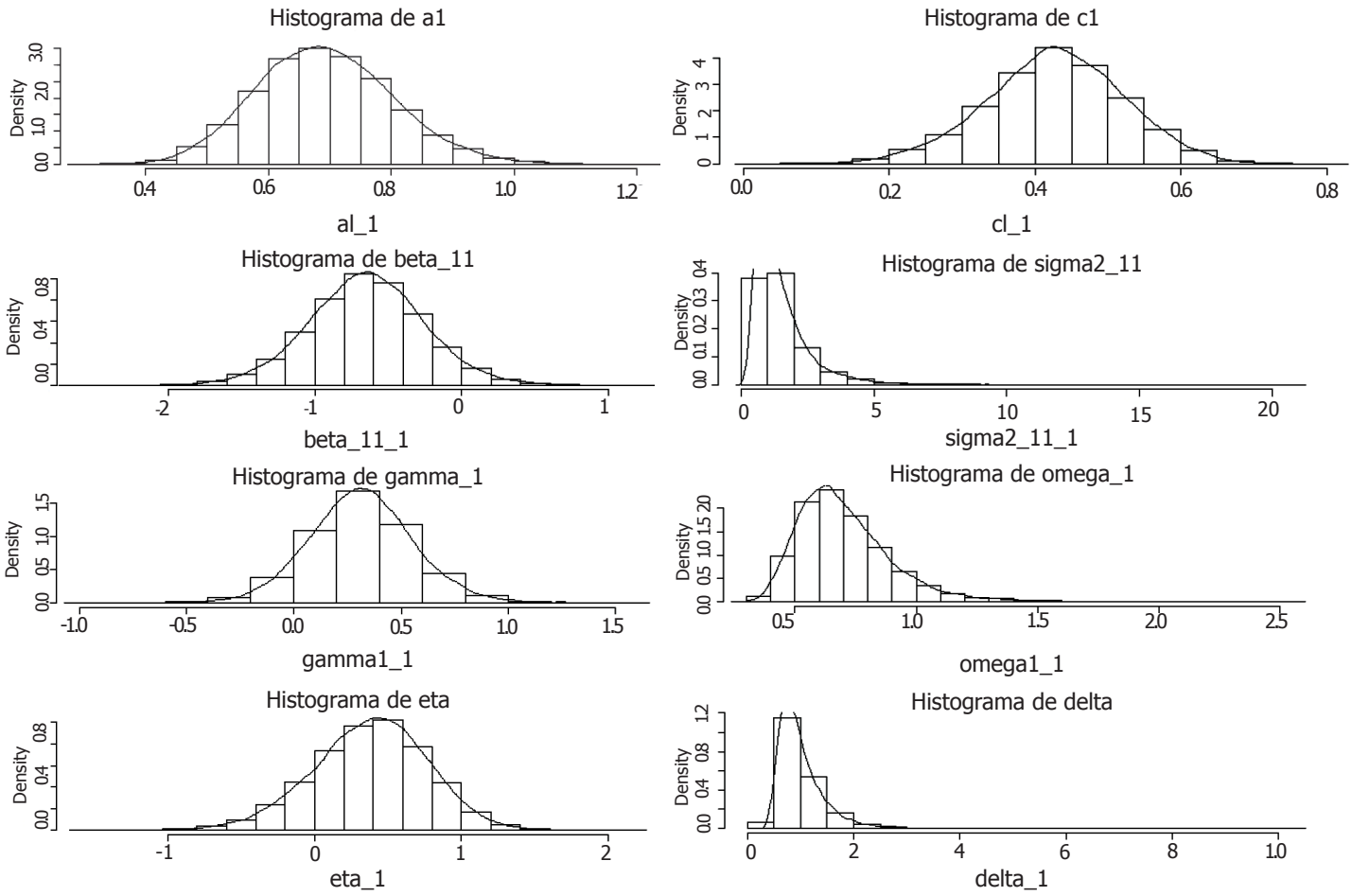

Figura 2. Histograma y estimador de la función de densidad para los ocho parámetros correspondientes a la figura 1.

La tabla 3 contiene las medias y desviaciones típicas de las 10 escuelas procedentes de cada región. La media resulta útil para identificar el nivel relativo de las distintas escuelas, detectando así posibles debilidades en el sistema escolar que puedan llevar a tomar medidas de mejora en aquellos casos en los que el nivel académico fuese claramente inferior. La desviación típica informa de cuánto varían entre sí los distintos estudiantes de una misma escuela, identificando así aquellas que tienen un nivel más o menos heterogéneo entre sus miembros.

De modo similar, la tabla 4 informa del nivel medio y la desviación típica en las tres regiones. Se aprecia que las medias difieren, siendo las regiones 1 y 3 las de mayor y menor nivel, respectivamente. En las desviaciones típicas no se aprecia un patrón tan claro. Esto sugiere la necesidad de indagar la razón por la cual existen estas diferencias en la educación que reciben los estudiantes. Las diferencias entre los promedios de las ciudades pueden analizarse mediante la inspección de los diagramas de

Tabla 3

Parámetros estimados para el modelo estructural de primer orden

	Región 1		Región 2		Región 3	
Escuela	β_{ij}	σ_{ij}	β_{ij}	σ_{ij}	β_{ij}	σ_{ij}
1	-0.66	1.53	0.21	2.01	1.19	3.02
2	-0.07	0.26	0.13	0.86	0.41	0.67
3	0.20	0.62	0.48	0.68	1.44	1.14
4	0.37	0.89	0.82	0.96	1.55	1.51
5	0.27	0.75	0.86	0.80	1.23	2.30
6	0.03	0.91	0.73	0.86	1.47	3.01
7	0.21	0.72	0.96	1.20	1.65	2.25
8	0.71	0.50	1.34	0.83	1.38	0.82
9	0.97	1.01	1.46	0.81	2.27	3.78
10	1.02	2.54	1.18	1.19	1.47	4.29

Nota: La tabla contiene la media y la desviación típica en cada escuela de cada región.

dispersión entre γ_k y ω_k en cada una de ellas. La figura 3 muestra el resultado, en el que se aprecia el incremento del promedio de una ciudad a otra, mientras que la dispersión permanece más estable.

Completando la jerarquía de niveles, la tabla 5 contiene la media y desviación típica para la muestra completa, juntando todos los datos. Más que por su valor numérico en sí mismo, estos valores sirven como punto de referencia para comparar los de cada región o escuela.

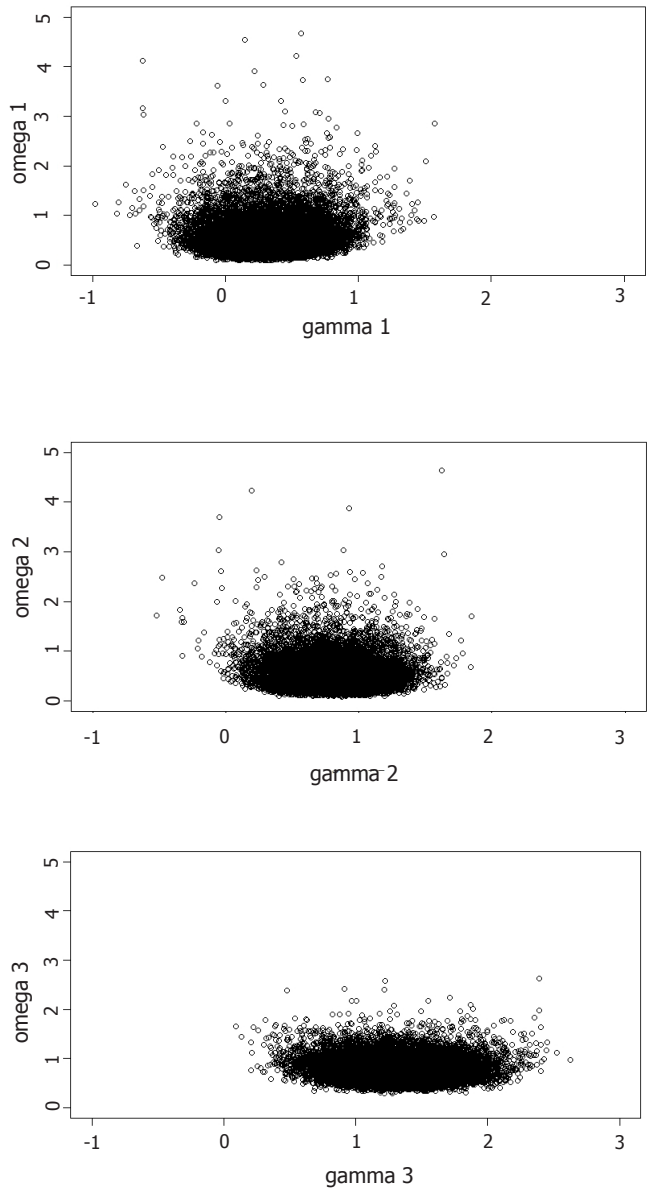

Figura 3. Diagrama de dispersión de la media (γ) y desviación típica (*ω*) de las puntuaciones en cada una de las tres regiones.

Tabla 4

Parámetros estimados para el modelo estructural de segundo orden

Parámetro	γ_{ij}	Se(γ_{ij})	ω_{ij}	Se(ω_{ij})
Región 1	0.31	0.25	0.54	0.33
Región 2	0.79	0.24	0.48	0.29
Región 3	1.32	0.30	0.63	0.41

Nota: La tabla contiene la media y la desviación típica en cada región.

Tabla 5

Parámetros estimados para el modelo estructural de tercer orden

Parámetro	Media	Se
η	0.38	0.39
δ	1.11	1.33

Nota: Datos para la muestra completa, juntando las escuelas de las tres regiones. La media de cada parámetro es el estimador esperado a posteriori. *Se* es el error típico a posteriori.

Por último, la tabla 6 muestra la correlación intraclase para cada escuela y región. Estas correlaciones son proporciones de varianza explicada, según se aprecia en las ecuaciones (8) y (9). La correlación intraclase de una escuela es alta cuando hay poca variabilidad dentro de dicha escuela, indicando que las puntuaciones dentro de la escuela están fuertemente relacionadas y las diferencias entre estos estudiantes y los del resto de la muestra se deben primordialmente al efecto de la región. Por el contrario, una correlación intraclase baja aparece cuando hay mucha variabilidad dentro de una escuela, lo que supone un efecto de la región más débil dentro de esa escuela.

Tabla 6

Correlaciones intraclase en cada escuela y región

Escuela	Región 1	Región 2	Región 3
1	.535	.496	.434
2	.839	.632	.714
3	.700	.674	.622
4	.622	.597	.549
5	.659	.639	.448
6	.619	.621	.385
7	.666	.546	.447
8	.744	.632	.667
9	.619	.645	.417
10	.451	.581	.389
	$\rho_2 = .627$	$\rho_2 = .646$	$\rho_2 = .599$

Nota: Las diez primeras filas de la tabla contienen la correlación intraclase (ρ_1) de cada escuela calculada según la fórmula (8). La fila inferior de la tabla contiene la correlación intraclase ρ_2 calculada según la fórmula (9).

Conclusiones

El presente trabajo aborda la problemática del desarrollo de modelos psicométricos para diseños de investigación multinivel que tienen como objetivo comparar el nivel medio de los sujetos pertenecientes a varios grupos organizados jerárquicamente. Tradicionalmente, esto se ha llevado a cabo mediante análisis de equivalencia/invarianza factorial en la medida (véase Revuelta y Ximénez, 2012; Ximénez y Revuelta, 2010). Los modelos multinivel suponen un paso más allá en los diseños multigrupo, que se utilizan para comparar los resultados de varios grupos organizados jerárquicamente.

En el artículo se ha expuesto un modelo multinivel basado en la Teoría de Respuesta al Ítem, TRI, y un procedimiento para obtener estimaciones en modelos jerárquicos de TRI mediante métodos Bayesianos. La estimación de modelos de TRI mediante métodos Bayesianos aplicados con simulación tiene una larga tradición que se origina en los trabajos de Albert (1992) y Albert y Chib (1993). Históricamente, los métodos Bayesianos se aplicaron inicialmente a los modelos dicotómicos (Patz y Junker, 1999a, 1999b) y posteriormente se extendieron a los multidimensionales (Béguin y Glas, 2001), politómicos (Revuelta, 2004, 2005), análisis factorial (Edwards, 2010; Ximénez, 2006) y a otros ámbitos aplicados (e.g., San Luis et al., 2011). Posteriormente, además de aplicarse a modelos preexistentes, la capacidad de los métodos Bayesianos para resolver problemas complejos de estimación de forma eficaz ha estimulado el desarrollo de modelos más sofisticados (Mellenbergh, 1994). Estos nuevos modelos se caracterizan por la inclusión de covariables para explicar el valor de determinados parámetros (de Boeck y Wilson, 2004), la inclusión de efectos aleatorios y las estructuras multinivel que permiten estructurar la muestra de una manera más rica.

Desde un punto de vista técnico, la estimación de los modelos de respuesta al ítem se basa en la distribución marginal de la matriz de respuestas (Baker y Kim, 2004). Cada efecto aleatorio implica que es necesario realizar una integral para calcular la distribución marginal de las respuestas. Por ello, en los modelos multidimensionales y en los de efectos aleatorios el cálculo de la distribución marginal implica una integral múltiple que no puede realizarse de forma eficaz por los métodos numéricos tradicionales.

Para solventar el problema de la estimación han surgido distintas soluciones. Una de ellas es el desarrollo de métodos de integración numérica más eficaces, como la cuadratura adaptativa de Gauss-Hermite (Schilling y Bock, 2005). Sin embargo, los métodos Bayesianos de simulación posiblemente sean la alternativa más extendida, porque su propia lógica se adapta de forma natural a los modelos de efectos aleatorios, en los que los parámetros pasan a ser considerados variables aleatorias. Además, los métodos Bayesianos permiten utilizar distribuciones a priori que facilitan la estimación cuando el modelo es complejo y la muestra resulta poco informativa para alguno de sus parámetros, lo cual puede suceder, por ejemplo, con muestras de tamaño medio o pequeño. De este modo se evita que unos pocos parámetros mal estimados impidan la estimación del modelo en su conjunto.

Un inconveniente de la estimación Bayesiana es que está poco automatizada en programas informáticos de uso general y por tanto requiere escribir un código informático para aplicarla, como se ha visto en este artículo. Además, estos programas requieren bastante tiempo de ejecución, lo que entorpece el proceso de probar, modificar y comparar varios modelos diferentes.

El presente trabajo se ha centrado en un tipo de modelo relativamente sencillo, el modelo dicotómico unidimensional, para poner el foco en el aspecto que se desea enfatizar: la estructura jerárquica de los datos y cómo el modelo proporciona información para cada nivel. Además de la exposición al modelo, se ha incluido un ejemplo con datos empíricos referidos a un test de matemáticas aplicado a estudiantes españoles, para ilustrar al lector sobre la aplicación de los procedimientos propuestos para el análisis jerárquico de cada nivel. Los resultados proporcionan información sobre cada estudiante, escuela y región.

En cuanto a las líneas de investigación futura, dos extensiones obvias del modelo consistirían en aplicarlo a datos politómicos, con más de dos categorías de respuesta, y multidimensionales, con dos o más rasgos latentes. Esto requeriría abordar nuevas problemáticas como la transformación de los niveles de rasgo mediante rotación ortogonal u oblicua, al igual que sucede en el análisis factorial exploratorio, o la realización de análisis confirmatorios.

Desde el punto de vista del investigador aplicado, la aplicación del modelo propuesto o sus extensiones puede realizarse a partir del código *OpenBUGS* o *Stan* de los apéndices A y B. Ambos programas utilizan el Gibbs-sampling. Sin embargo, *Stan* se basa en la denominada dinámica Hamiltoniana (Neal, 2011), que constituye un método más moderno y eficaz para tomar muestras aleatorias que el algoritmo implementado en *OpenBUGS* (Gilks, et al., 1996). Además, el programa *Stan* se ha elaborado utilizando el lenguaje informático C++, que proporciona programas de ejecución rápida. Debido a su mayor novedad, existe menos bibliografía y ejemplos disponibles sobre su manejo, aunque cabe esperar que su popularidad aumente en el futuro y llegue a tener un papel más dominante dentro del campo de la computación Bayesiana.

Referencias

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251-269. doi: 10.3102/10769986017003251

Albert, J. H. y Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669-679. doi:10.1080/01621459.1993.10476321

Alwin, D. F. y Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. En D. D. Jackson y E. P. Borgatta (Eds.), *Factor analysis and measurement in sociological research: a multidimensional perspective* (pp. 249-280). Beverly Hills, CA: Sage.

Baker, F. B. y Kim, S. H. (2004). *Item response theory. Parameter estimation techniques*. Boca Ratón, FL: Chapman & Hall/CRC.

Béguin, A. A. y Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66*, 471-488. doi: 10.1007/BF02296195

Brooks, S. y Gelman, A. (1998). General methods for monitoring convergence of iterative simulation. *Journal of computational and Graphical Statistics*, *7*, 434-455. doi:10.1080/10618600.1998.10474787

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29.

de Boeck, P. y Wilson, M. (2004). *Explanatory item response models. A generalized linear and non-linear approach*. New York: Springer.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika, 75*, 474-497. doi: 10.1007/s11336-010-9161-9

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous items. *British Journal of Mathematical and Statistical Psychology, 58*, 145-172. doi: 10.1348/000711005X38951

Fox, J. P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software, 20*, 1-16. Recuperado de http://doc.utwente.nl/59662/5/Fox07multilevel.pdf

Fox, J. P. (2010). *Bayesian item response theory*. New York: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. y Rubin, D. B. (2014). *Bayesian data analysis*. Third edition. Boca Ratón, FL: Taylon & Francis.

Gelman, A. y Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-511. Recuperado de: http://www.jstor.org/stable/2246093

Gilks, W. R., Richardson, S. y Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Boca Ratón, FL: Chapman & Hall/CRC.

Goldstein, H. (2003). *Multilevel statistical models*. Third edition. Londres: Hodder Arnold.

Goldstein, H. (2004). International comparison of student attainment. Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice, 11*, 319-330. doi: 10.1080/0969594042000304618

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

Hoffman, M. D. y Gelman, A. (2011). The No-U-turn sampler: Adaptively setting paths lengths in Hamiltonian Monte Carlo. *arXiv:1111.4246* [*stat.CO*]. Recuperado de http://arxiv.org/abs/1111.4246

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426. doi: 10.1007/BF02291366

Martínez Arias, R., Hernández, M. V. y Hernández, M. J. (2006). *Psicometría*. Madrid: Alianza.

McCulloch, C. E., Searle. S. R. y Neuhaus, J. M. (2008). *Generalized, linear and mixed models*. New York: Wiley.

Mellenbergh, G. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 200-307. doi: 10.1037/0033-2909.115.2.300

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543. doi: 10.1007/BF02294825

Neal, R. (2011). MCMC using Hamiltonian dynamics. En S. Brooks, A. Gelman, G. L. Jones y X-L. Meng, (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 113-162). Boca Ratón, FL: Chapman & Hall/CRC.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. New York: Wiley.

Osterlind, S. J. y Everson, H. T. (2009). *Differential item functioning*. Thousand Oaks, CA: Sage Publishing.

Patz, R. J. y Junker, B. W. (1999a). A straightforward approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics, 24*, 146-178. doi: 10.3102/10769986024002146

Patz, R. J. y Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366. doi: 10.3102/10769986024004342

Revuelta, J. (2004). Analysis of distractor difficulty in multiple-choice items. *Psychometrika, 69*, 217-234. doi: 10.1007/BF02295941

Revuelta, J. (2005). An item response model for nominal data based on the rising selection ratios criterion. *Psychometrika, 70*, 305-324. doi: 10.1007/s11336-002-0975-y

Revuelta, J. y Ximénez, C. (2012). Mean structure analysis from an IRT approach: An application in the context of organizational psychology. *Psicothema, 24*, 653-660. Recuperado de http://www.unioviedo.es/reunido/index.php/PST/article/viewFile/9718/9462

San Luis, C., Cañadas, G. A., Cantero, J., Lozano, L. M., de la Fuente, E. I. y Lozano, T. (2011). Applicability of the Bayesian methodology to the study of low incidence diseases: Example of child anxiety. *Revista de Psicopatología y Psicología Clínica*, *16*, 61-66. Recuperado de http://e-spacio.uned.es/revistasuned/index.php/RPPC/article/view/10351

Schilling, S. y Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533-555. doi: 10.1007/s11336-003-1141-x

Spiegelhalter, D., Thomas, A., Best, N. y Lunn, D. (2003). *WinBUGS User manual. Version 1.4.* MRC Biostatistics Unit, Institute of Public Health, Robinson Way: Cambridge CB2 2SR, UK.

Spiegelhalter, D., Thomas, A., Best, N. y Lunn, D. (2014). *OpenBUGS User manual.* MRC Biostatistics Unit, Institute of Public Health, Robinson Way: Cambridge CB2 2SR, UK. Recuperado de: http://www.openbugs.net/w/FrontPage

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling*, *13*, 587-614. doi: 10.1207/s15328007sem1304_5

Ximénez, C. y Revuelta, J. (2010). Factorial invariance in a repeated measures design. An application to the study of person-organization fit. *Spanish Journal of Psychology, 13*, 485-493. doi:10.1017/S1138741600004030

Fecha de recepción: 8 de abril de 2014
Fecha de aceptación: 3 de junio de 2014

Apéndice A

Estimación mediante el programa *OpenBUGS*

El modo más conveniente de realizar la estimación Bayesiana con el programa *OpenBUGS* es utilizando el denominado modo interactivo, o modo *batch*, que permite la ejecución automática del programa a partir de varios archivos de datos y sintaxis. De este modo, es posible replicar fácilmente la estimación modificando los datos o códigos del programa y sin tener que utilizar menús. Para elaborar el código y ejecutar *OpenBUGS* en modo interactivo se han seguido los pasos descritos en los apéndices B y C del libro de Ntzoufras (2009).

La estimación interactiva del ejemplo de este artículo requiere cuatro tipos de archivos, que suponemos están almacenados en la carpeta `C:\BayesIRT` del ordenador. Los archivos son:

1. Archivo script (l_scale_script.odc). Es el archivo que dirige todo el proceso de estimación. Desde ahí se carga el archivo de datos, el archivo de instrucciones donde está programado el modelo estadístico y los archivos con los valores iniciales para comenzar la simulación. También indica qué variables deben almacenarse como resultado y el nombre del archivo de resultados. En el presente análisis se ha utilizado el programa *OpenBUGS*, que es una versión más actual de *WinBUGS*. La sintaxis de los archivos script de ambos programas guarda bastantes diferencias. El libro de Ntzoufras (2009) se basa en *WinBUGS*, mientras que en este apartado se utiliza la sintaxis de *OpenBUGS* que aparece descrita en el propio programa accediendo al menú `Manuals > OpenBUGS User Manual`.

```
# Crea el archive log donde se guardarán los resultados
modelDisplay('log')

# Comprobar la sintaxis del código
modelCheck('C:/BayesIRT/l_scale_code.odc')

# Cargar los datos y constantes
modelData('C:/BayesIRT/l_scale_data.odc')
modelData('C:/BayesIRT/l_scale_constants.odc')

# Especificar cuatro cadenas de Markov
modelCompile(4)

# Valores iniciales para cada cadena
modelInits('C:/BayesIRT/l_scale_initial_1.odc', 1)
modelInits('C:/BayesIRT/l_scale_initial_2.odc', 2)
modelInits('C:/BayesIRT/l_scale_initial_3.odc', 3)
modelInits('C:/BayesIRT/l_scale_initial_4.odc', 4)

# Valores aleatorios iniciales para los parámetros sin especificar
modelGenInits()

# Tomar muestras para el periodo de arranque inicial
modelUpdate(20000)

# Parámetros que se van a archivar como resultado
samplesSet('a')
samplesSet('c')
samplesSet('beta')
samplesSet('gamma')
samplesSet('eta')
samplesSet('sigma2')
samplesSet('omega2')
samplesSet('delta2')
```

```
# Tomar muestras, pasado el periodo de arranque
modelUpdate(20000,2)

# Incluir estadísticos y gráficos en el archive de resultados
samplesStats('*')
samplesDensity('*')
samplesHistory('*')

# Guardar parámetros en archivos de texto con formato CODA
samplesCoda('*', 'C:/BayesIRT/l_scale_ac')
modelSaveLog('C:/BayesIRT/l_scale_log.odc')

# Cerrar el programa
modelQuit('yes')
```

2. Archivos de datos (l_scale_data.odc y l_scale_constants.odc). Son archivos que deben crearse desde dentro del programa *OpenBUGS*, que los guarda en su formato propio. El archivo l_scale_data.odc contiene la matriz de respuestas dicotómicas y las variables R y S, que indican la región y escuela. La primera línea del archivo contiene los nombres de las variables, que deben coincidir con los nombres dados en el código *OpenBUGS* del archivo l_scale_code.odc, descrito en el punto 3. El archivo l_scale_data.odc debe acabar con la palabra clave END seguida de una línea en blanco. En el ejemplo, el archivo l_scale_data.odc tiene el aspecto:

```
 X[,1]  X[,2]  X[,3]  X[,4] X[,5] X[,6] X[,7] X[,8] X[,9] X[,10] X[,11] X[,12] R[] S[]
   1      1      1      0     1     1     1     0     1      1      1      1     1   1
   0      0      0      0     0     1     1     1     0      0      1      0     1   2
   0      0      0      0     0     0     0     1     1      1      0      0     1   1

(varias decenas de líneas de datos se han omitido aquí por brevedad)

   1      1      1      1     0     1     0     0     0      0      1      1     3   10
   1      1      1      1     1     1     0     1     1      0      1      1     3   3
   1      1      0      1     0     1     0     1     1      1      0      1     3   1
END
```

El archivo l_scale_constants.odc simplemente contiene una serie de constantes que luego serán utilizadas por el archivo de código. Estas constantes, por orden, son el número total de sujetos, el número de ítems del test, el número de regiones y el de escuelas. El archivo l_scale_constants.odc sólo tiene una línea de contenido, que es:

```
list(N=1000, K=12, G=3, L=10)
```

3. Archivo de código. Contiene el modelo estadístico programado en lenguaje *OpenBUGS* (l_scale_code.odc).

```
model
{
    for (i in 1:N) {
      for (k in 1:K-1) {
       p[i,k] <- exp(a[k]*theta[i] + c[k])/
              (1+exp(a[k]*theta[i] + c[k]))
             X[i,k] ~ dbern(p[i,k])
       }

       p[i,K] <- exp(theta[i])/(1+exp(theta[i]))
       X[i,K] ~ dbern(p[i,K])
```

```
     theta[i] ~ dnorm(beta[R[i], S[i]],
                 tau_sigma2[R[i], S[i]])
   }

   for (k in 1:K-1) {
      a[k] ~ dgamma(0.25, 0.25)
      c[k] ~ dnorm(0, 4)
   }

   for(g in 1:G){                  # Región
        gamma[g] ~ dnorm(eta, tau_delta2)
        tau_omega2[g] ~ dgamma(1, 1)
        omega2[g] <- 1/tau_omega2[g]

        for(l in 1:L){               # Escuela
            beta[g,l] ~ dnorm(gamma[g], tau_omega2[g])
       tau_sigma2[g,l] ~ dgamma(1, 1)
       sigma2[g,l] <- 1/tau_sigma2[g,l]
        }
   }

   eta ~ dnorm(0, 4)
   tau_delta2 ~ dgamma(1, 1)
   delta2 <- 1/tau_delta2
}
```

4. Archivos de valores iniciales para la estimación (l_scale_initial_1.odc, l_scale_initial_2.odc, l_scale_initial_3.odc y l_scale_initial_4.odc). Son valores iniciales arbitrarios asignados a los parámetros *a* y *c*. Determinan el punto de inicio de las cadenas, y podían haberse empleado otros valores distintos. El contenido de cada uno de estos archivos son dos líneas de código. El archivo l_scale_initial_1.odc contiene:

```
list(a=c( 1.1,   1.2,   1.3,  0.9,  0.7,   0.8,   1.0,   1.1,   0.8,   1.4,   0.9))
list(c=c( -1,     0,     1,    -2,  -0.5, -0.8,  0.2,   1.4,   -0.9,   -1,     0))
```

Los otros tres archivos de valores iniciales son:

```
list(a=c( 0.9,   1.2,   0.7,  1.4,  1.1,   1.2,   1.3,   0.9,   0.7,   0.8,   1.0))
list(c=c( 0.5,    1,    -2,  -0.7,  -1,    0,     1,    -2,   -0.5,  -0.8,  0.2))

list(a=c( 1.4,   0.8,   1.1,  1.2,  0.9,   1.2,   0.7,   1.4,   1.1,   1.2,   1.3))
list(c=c(-0.5,   0.1,    2,  -1.7,  0.5,    1,    -2,   -0.7,   -1,     0,     1))

list(a=c( 0.8,   1.4,   1.0,  1.3,  1.4,   0.8,   1.1,   1.2,   0.9,   1.2,   0.7))
list(c=c( 0.9,  -1.2,   0.5,  0.7, -0.5,   0.1,    2,   -1.7,   0.5,    1,    -2))
```

Como resultado, el programa genera dos tipos de archivos:

1. Archivos en formato texto, que pueden ser manipulados fácilmente por otros programas para realizar otros análisis. Una forma sencilla de manipular estos archivos es utilizando el paquete CODA mediante el lenguaje de programación estadística R.

2. Archivo de resultados en formato *OpenBUGS*. Solo puede abrirse con este programa, y contiene diversas tablas de estadísticos descriptivos y gráficos sobre el resultado de la simulación.

Apéndice B

Estimación mediante el programa *Stan*

El programa *Stan* permite aplicar los métodos Bayesianos utilizando algoritmos de estimación más eficaces que los incluidos en *OpenBUGS*. En el apéndice C del libro de Gelman et al. (2014) aparece una introducción a *Stan* que incluye indicaciones sobre la instalación y varios ejemplos.

En este trabajo se ha utilizado el programa *Stan* dentro del entorno de programación R, utilizando el paquete *RStan*. Las instrucciones de instalación y el manual del usuario pueden obtenerse en la página web http://mc-stan.org/. Para ejecutar *RStan* son necesarios tres archivos: un archivo de datos, un archivo de código R y un archivo de código *Stan*.

1. El archivo de datos debe tener formato texto, y su contenido es una matriz de datos con sujetos en las filas y variables en las columnas. Las primeras tres filas del archivo de datos del ejemplo son:

```
1   1   1   0   1   1   1   0   1   1   1   1   1   1
0   0   0   0   0   1   1   1   0   0   1   0   1   2
0   0   0   0   0   0   0   1   1   1   0   0   1   1
```

   Puede verse que los datos son los mismos que los que contiene el archivo l_scale_data.odc descrito en el apéndice A, pero eliminando cualquier elemento que no sean estrictamente los datos de la muestra. Las primeras 12 columnas son las respuestas a los ítems, donde 0 y 1 indica error y acierto, y la respuesta perdida se ha codificado con el valor 2. Las columnas 13 y 14 son la región y la escuela, respectivamente. A este archivo se le ha dado el nombre l_scale_data.dat.

2. Archivo de código R lleva por nombre l_scale_R_script.r y es el archivo que dirige todo el proceso. Lee el archivo de datos y llama al programa *Stan* pasándole el archivo de datos y otros códigos necesarios para la estimación. Debe ejecutarse desde dentro del entorno de programación estadística R. Su contenido es:

```
library(rstan)

#Leer los datos y los prepara para que Stan los utilice
Y <- as.matrix(read.table("l_scale_data.dat", header=F))

nc <- ncol(Y)
X=Y[,1:(nc-2)]
R=Y[,nc-1]
S=Y[,nc]

N <- nrow(X)     # Number of examinees
J <- ncol(X)     # Number of items
G <- max(R)      # Number of regions
L <- max(S)      # Number of schools

# Prepara un objeto con todos los datos que va a requerir Stan
l_scale_dat <- list(N=N, J=J, G=G, L=L, X=X, R=R, S=S)

# Número de iteraciones para la estimación
iter <- 40000

# Llama a Stan y le pasa los datos
l_scale_fit <- stan(file = 'l_scale_code.stan', data = l_scale_dat, iter=iter, verbose =
FALSE)

# Muestra los resultados por pantalla
print(l_scale_fit, pars = c("a", "c", "beta", "sigma", "ro_1", "gamma", "omega", "ro_2",
"eta", "delta"))
```

3 . Archivo de código *Stan*. Es el archivo en el que está programado el modelo estadístico. Lleva por nombre `l_scale_code.stan` y debe tener formato texto. El contenido del archivo para este ejemplo es:

```
// Datos para el modelo estadístico, en el mismo orden
// que los datos que se le pasan al programa desde el
// archivo l_scale_R_script.r

data {
  int<lower=0> N;     // number of examinees
  int<lower=0> J;     // number of items
  int<lower=0> G;     // number of regions
  int<lower=0> L;     // number of schools
  int X[N,J];         // responses
  int R[N];           // region
  int S[N];           // school
}

// Lista de parámetros del modelo estadístico
parameters {
  real theta[N];
  real<lower=0> a[J-1];
  real c[J-1];
  real beta[G,L];
  real gamma[G];
  real eta ;
  real<lower=0.0001> delta2 ;
  real<lower=0.0001> omega2[G];
  real<lower=0.0001> sigma2[G,L];
}

// Desviaciones típicas, calculadas como la raíz de las
// varianzas, y correlaciones intraclase
transformed parameters {
  real<lower=0> delta ;
  real<lower=0> omega[G];
  real<lower=0> sigma[G,L];

  real<lower=0> ro_1[G,L]; // Intraclass corr., escuelas
  real<lower=0> ro_2[G]; // Intraclass corr., regiones

  for(g in 1:G){                 // Region
      omega[g] <- sqrt(omega2[g]);
      ro_2[g] <- (delta2) / (delta2 + omega2[g]);
      for(l in 1:L){             // Escuela
          sigma[g,l] <- sqrt(sigma2[g,l]);
          ro_1[g,l] <- (delta2 + omega2[g]) /
                       (delta2 + omega2[g] + sigma2[g,l]);
      }
  }
  delta <- sqrt(delta2);
}

// Modelo estadístico de TRI multinivel
model {
    for(i in 1:N){
        // Muestrear theta, eq. (3) del artículo
        theta[i] ~ normal(beta[R[i], S[i]], sigma[R[i], S[i]]);
```

```
        for(j in 1:J-1){
             if(X[i,j] < 2){
                // Eq. (1) del artículo
                 X[i,j] ~ bernoulli_logit(a[j]*theta[i] + c[j]);
             }
        }

        if(X[i,J] < 2){ // Si X > 2 se trata de un dato perdido
            X[i,J] ~ bernoulli_logit(theta[i]);
        }
    }
    a ~ gamma(0.25, 0.25);
    c ~ normal(0, 1);

    for(g in 1:G){                  // Region
       // Eq. (5) del artículo
        gamma[g] ~ normal(eta, delta);
        omega2[g] ~ gamma(1, 1);
        for(l in 1:L){               // Escuela
           // Eq. (4) del artículo
            beta[g,l] ~ normal(gamma[g], omega[g]);
            sigma2[g,l] ~ gamma(1, 1);
        }
    }
    eta ~ normal(0, 1);
    delta2 ~ gamma(1, 1);
}
```