UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Estadística en la evaluación de la educación
Junio 2016, Vol. 9, No. 1, pp. 127–146

Algunos retos metodológicos del modelamiento bayesiano de teoría de respuesta al ítem: la calificación de pruebas estandarizadas[1]

Some methodological challenges in bayesian item response modeling: the assessment of standardized tests

Andrés Gutiérrez[a]
agutierrez@icfes.gov.co

Diego Fernando Lemus[b]
dlemus@icfes.gov.co

William Fernando Acero[c]
wacero@contratista.icfes.gov.co

Resumen

El Ministerio de Educación Nacional de Colombia (MEN) se preocupa cada vez más por la mejora continua en los procesos de enseñanza de los docentes que enseñan una lengua extranjera, y ha encargado la evaluación de dicho proceso al Instituto Colombiano para la Evaluación de la Educación (Icfes), con el fin de evaluar los conocimientos de dichos docentes para que los mismos cumplan con los estándares establecidos en el Marco Común Europeo de Referencia (MCER). La reducción del error de estimación resulta ser un objetivo fundamental en la calificación de las pruebas implementadas por el Icfes, con el fin de mejorar la precisión al momento de evaluar parámetros como la dificultad del examen y el desempeño de quienes lo presentan. En este artículo se encontrará una breve comparación entre la metodología clásica del modelo de Rasch y la metodología bayesiana del mismo, comparando el comportamiento y conservación de los supuestos clásicos en la segunda metodología. Además se muestra cómo la metodología bayesiana reproduce una disminución en el error de estimación de los parámetros de habilidad y dificultad frente a la metodología clásica.

***Palabras clave*:** modelo de Rasch, metodología bayesiana, error de estimación .

[1]Gutiérrez, A., Lemus, D., H., Acero, W. (2016) Algunos retos metodológicos del modelamiento bayesiano de teoría de respuesta al ítem: la calificación de pruebas estandarizadas. *Comunicaciones en Estadística*, **9**(1), 127-146.

[a]Director de Evaluación, Icfes
[b]Subdirector de Diseño de Instrumentos, Dirección de Evaluación, Icfes
[c]Estadístico, Dirección de Evaluación, Icfes

Abstract

The Ministerio de Educación Nacional de Colombia (MEN), is increasingly concerned about the continuous improvement on teaching processes of the foreign languages teachers and has commissioned the assessment of this process to the Instituto Colombiano para la Evaluación de la Educación (Icfes), in order to assess the knowledge of those teachers and ensure that they conforms to the quality standards of the Common European Framework of Reference for Languages (CEFR). The reduction of the estimation errors results as a key objective in the rating process of the tests That are Implemented by the Icfes, in order to improve increase accuracy for Evaluating parameters: such as the difficulty of the test and the performance of Those Who present the test. In a brief comparison este documento Between classical and Bayesian methodology for Rasch model is provided, esta comparison contains the behavior and upkeep of the classic assumptions for the second methodology. Furthermore the reduction of the errors for the Ability and difficulty parameters in the Bayesian framework is presented.

***Keywords*:** Rasch model , Bayesian Methodology, Estimation error.

1. Introducción

En la sociedad actual resulta de gran interés evaluar constantemente el aprendizaje de los individuos. Para tal fin, las instituciones educativas, industrias e incluso el gobierno nacional aplican exámenes para determinar el avance del proceso cognitivo del sistema educativo. Bajo este escenario resulta de vital importancia desarrollar examenes estandarizados que se acerquen cada vez más a lo que se quiere evaluar. El Icfes, tiene como objeto fundamental ofrecer el servicio de evaluación de la educación en todos sus niveles y adelantar investigación sobre los factores que inciden en la calidad educativa, con la finalidad de ofrecer información para mejorar la calidad de la educación.

El MEN con el fin de definir un sistema de evaluación sólido y coherente, ha establecido líneas claras para la identificación de las necesidades de formación de los docentes, la formulación de planes de capacitación, y en general, el monitoreo cercano de los procesos de enseñanza y aprendizaje del inglés en el país. Bajo esa premisa, el MEN adoptó el Marco Común Europeo de Referencia para el aprendizaje, la enseñanza y la evaluación del inglés como lengua extranjera desde el 2004 e implementó el Programa Nacional de Bilingüismo como estrategia orientada a elevar la competencia en idioma inglés en el ámbito nacional. Desde 2008, el MEN ha venido implementando pruebas diagnósticas de nivel de inglés para docentes, con una frecuencia anual y con el objetivo de contar con información real y actualizada sobre los niveles de inglés de los docentes del sector oficial colombiano.

La experiencia adquirida durante los 40 años de existencia del Icfes en el diseño y ejecución de diversos tipos de evaluación le ha permitido desarrollar la capacidad técnica y operativa para realizar otras evaluaciones que le sean encargadas por

entidades públicas o privadas y derivar de ellas ingresos, conforme a lo establecido en la Ley 635 de 2000. Por tal motivo, desde 2015, el Icfes bajo la dirección de MEN evaluó el nivel de uso de inglés de los docentes licenciados en este lenguaje, según el Marco Común Europeo de Referencia para las lenguas: aprendizaje, enseñanza y evaluación -MCER. La prueba de inglés solicitada evaluó tres habilidades del lenguaje a través de tres componentes como se describen a continuación:

- Componente de escucha: este componente cuenta con 30 preguntas para ser respondidas en 40 minutos y permite clasificar a los evaluados en uno de los siguientes niveles del Marco Común Europeo a través de cinco distintas partes: A1 o inferior, A2, B1, B2 o superior.
- Componente de lectura: este componente cuenta con 45 preguntas para ser respondidas en 60 minutos y permite clasificar a los evaluados en uno de los siguientes niveles del Marco Común Europeo: Inferior a A1, A1, A2, B1, B2 o superior.
- Componente de escritura: este componente contiene 2 partes para ser respondidas en 45 minutos. Los evaluados deben elaborar dos escritos: uno corto y otro extenso atendiendo a un contexto determinado y respondiendo a unas condiciones dadas. Las dos partes permiten clasificar a los evaluados en uno de los siguientes niveles del Marco Común Europeo (A1 o inferior, A2, B1, B2 o superior.

En este artículo se presentan los resultados obtenidos en el proceso de calificación del componente de escucha de la aplicación 2015 de la prueba de inglés (implementada a un grupo de 8950 docentes licenciados en inglés) utilizando tanto el enfoque clásico como el enfoque bayesiano de la teoría de respuesta al ítem. Este proyecto de investigación está enmarcado dentro de uno de los objetivos de la Dirección de Evaluación del Icfes que consiste en mejorar los procesos de calificación que se tienen actualmente, a través de la implementación de nuevas metodologías que permitan obtener estimaciones de las dificultades de los ítems y de las habilidades de las personas, más precisas y con menor error. Para ello, la estadística bayesiana da un valor agregado a los procesos actuales puesto que, al poseer información previa de las pruebas, se puede utilizar para obtener mejores resultados en los procesos de calificación. Esta metodología podría ser adoptada en un futuro como parte del procesamiento estadístico de las pruebas.

El documento tiene la siguiente estructura: en la siguiente sección se presenta el marco metodológico considerado para el desarrollo del proyecto de investigación; en la sección 3 se presenta la metodología propuesta y en la cuarta sección los resultados obtenidos. Finalmente se presentan las conclusiones del estudio.

2. Marco metodológico

Fox (2010) afirma que los modelos para la teoría de respuesta al ítem (TRI) fueron

desarrollados entre 1970 y 1980, principalmente desde el campo de la psicometría, intentando evaluar rasgos latentes (como la habilidad de la persona), para así obtener mayor certeza de la conclusiones que podían sacar de sus estudios, algunos de los modelos que surgieron en esta época son: modelo de Rasch, modelo de un parámetro (modelo de 1 PL), de dos parámetros (modelo de 2 PL), entre otros. En Sinharay (2003) se manifiesta que el limitante computacional era un factor clave, ya que los métodos de estimación de estos modelos requieren una carga computacional importante, que no podía ser soportada por los desarrollos de la época por la creciente complejidad de las situaciones en las que se recogen los datos de respuesta plantea nuevos inconvenientes. Trabajos como los de Mislevy (1986), Rigdon & Tsutakawa (1983) y Swaminathan & Gifford (1982) presentan extensiones bayesianas de los modelos de respuesta al ítem tradicionales.

Según Ayala (2008), el modelo logístico de un parámetro propuesto por Rasch (1980) es uno de los más usados en la teoría de respuesta al ítem (TRI) [1], ya que sus costos computacionales son bajos y produce muy buenos resultados. Bajo el modelo logístico de un parámetro, la probabilidad de responder correctamente un ítem se define matemáticamente como:

$$P(Y_{ik} = 1|\theta_i, b_k) = \frac{e^{(\theta_i - b_k)}}{1 + e^{(\theta_i - b_k)}} = (1 + e^{(b_k - \theta_i)})^{-1} \tag{1}$$

Donde θ_i es la habilidad del individuo i, $i = 1, \cdots, n$, y b_k es la dificultad del ítem k, $k = 1, \cdots, J$. En otras palabras, bajo el modelo de Rasch la probabilidad de que un individuo conteste correctamente depende de su habilidad y de la dificultad del ítem. Como es de esperarse, a medida que la habilidad de una persona aumenta, la probabilidad de responder correctamente un ítem de dificultad b_k también, como se observa en la figura 1.

El enfoque bayesiano de la TRI hace uso de métodos computacionales mejorados para el modelamiento de la naturaleza discreta de los datos en la teoría de respuesta al ítem y, así generar un nuevo punto de vista más flexible que se ocupa de las relaciones con los datos de nivel superior, donde supuestos de distribución estándar no se aplican. Un elemento clave fue el desarrollo de los métodos MCMC (Monte Carlo Markov Chains) y su simplicidad para la estimación conjunta a pesar del aumento en la complejidad del modelo. Problemas específicos relacionados con el modelado de datos de respuesta hacen ciertos métodos bayesianos muy útil.

[1] Se entiende como ítem a la pregunta de un *test*

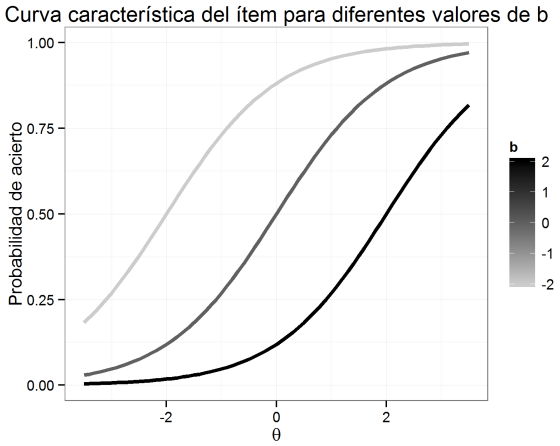

Figura 1: *Curva característica del ítem para diferentes* b_k*. Fuente: elaboración propia.*

Según Fox (2010) la habilidad del individuo θ y la dificultad de un ítem b_k se consideran variables aleatorias en la TRI bayesiana y estos parámetros una distribución *prior* que refleja la incertidumbre acerca de los verdaderos valores de estos parámetros antes de haber obtenido los datos. Los modelos de respuesta al ítem discutidos para los datos observados describen el proceso de generación de datos como una función de parámetros desconocidos en los que se conocen como modelos de probabilidad. Esta es la parte del modelo que presenta la densidad de los datos condicionales en los parámetros del modelo. Por lo anterior, se tienen dos pasos en el proceso de modelamiento: en el primero se realiza la especificación de una distribución *a priori* y en el segundo se realiza la especificación de un modelo de probabilidad. En este caso, podemos asumir distribuciones prior para los parámetros de la siguiente manera:

$$\theta \sim Normal(\mu_1, \tau_1^2) \tag{2}$$

$$b \sim Normal(\mu_2, \tau_2^2) \tag{3}$$

Donde el vector de hiperparámetros está definido por $(\mu_1, \tau_1^2, \mu_2, \tau_2^2)$. En particular estas distribuciones *prior* que se proponen son viables ya que tanto el parámetro θ como el parámetro b se encuentran en el intervalo $(-\infty, \infty)$, además cada $Y_{ik} \sim Bernoulli(p_{ik})$, donde p_{ik} es la probabilidad de responder acertadamente el $k-esimo$ ítem. Suponga que se tienen N realizaciones independientes (N individuos presentando una prueba) para cada uno de los k ítems de un examen, los cuales a su vez miden un único constructo de manera independiente. Por lo tanto la función de verosimilitud de los datos está dada por:

$$f(Y, b, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} p_{ik}^{y_{ik}} (1 - p_{ik})^{y_{ik}} \tag{4}$$

Además se tiene que cada p_{ik} está dado por la ecuación 1 y reemplazando 1 en 4 se obtiene:

$$f(Y, b, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} e^{(\theta_i - b_k) y_{ik}} (1 - e^{(\theta_i - b_k)})^{y_{ik}} \tag{5}$$

Teniendo en cuenta que la estimación de la habilidad se ve afectada por la estimación de la dificultad, tal como se muestra en la verosimilitud, es pertinente usar los métodos de estimación bayesiana para generar inferencias estadísticas más confiables sobre las dificultades de los ítems (Fox 2010). Para tal fin, puede considerarse la información obtenida en aplicaciones anteriores de una misma prueba; después de observar los datos, se combina la información de la distribución *a priori* con la información obtenida en la presente aplicación para generar densidades posterior que permitan hacer inferencia directa sobre los parámetros de interés. La flexibilidad en la definición de los modelos de TRI para los parámetros de interés hace posible para manejar, por ejemplo, diseños de muestreo más complejos que comprenden estructuras de dependencia complejas y es uno de los puntos fuertes del enfoque bayesiano.

3. Procedimiento propuesto

Como se mencionó en la sección anterior, en el enfoque bayesiano, los parámetros del modelo son variables aleatorias y tienen distribuciones *apriori* que reflejan la incertidumbre acerca de los verdaderos valores de dichos parámetros antes de haber observado los datos. En este sentido hay dos cosas fundamentales a tener en cuenta: la primera, la especificación de las distribuciones *apriori*; y la segunda, la función de verosimilitud del modelo, para que sea posible la combinación de los dos métodos de estimación, esto es, bayesiano y clásico. En este sentido, la inferencia bayesiana sobre los parámetros se realizan bajo las distribuciones condicionales de las densidades *posterior*.

Por otra parte, recordando el teorema de Bayes, se asume que la respuesta de los datos viene dada por una variable latente θ y, por tanto, $p(\theta)$ representa la información diponible *prior* acerca de estos datos y $p(y|\theta)$ hace referencia a la información observada de los datos; bajo este esquema es posible construir la siguiente relación:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{6}$$

Donde $p(\theta|y)$ es la distribución *posterior*.

3.1. Ejemplo A

Inicialmente se considera un ejemplo propuesto por Fox (2010), en el cual se supone que un estudiante con habilidad θ tiene el siguiente vector de respuestas

dicotomicas $\mathbf{y} = (1, 1, 0, 0, 0)^t$, donde 1 indica que el estudiante responde de manera correcta al ítem y 0 cuando no. El objetivo del ejemplo consiste en estimar la distribución *posterior* para el parámetro θ.

Un caso particular consiste en asumir que todos los ítems tienen la misma dificultad (por ejemplo, ítems de dificultad media donde $b_k = 0$). Bajo este escenario, la versión probit del modelo de Rasch donde $P(Y_k = 1|\theta) = \Phi(\theta)$ define la probabilidad de responder correctamente el *k-ésimo* ítem. Entonces para θ se puede asumir una distribución *prior* uniforme continua en el intervalo [-3,3] tal que $0.001 < \Phi(\theta) < 0.998$. La función de verosimilitud $p(y|\theta)$ está dada por:

$$p(y|\theta) = \Phi(\theta)^2(1 - \Phi(\theta))^3 \tag{7}$$

Así, multiplicando por la distribución prior $p(\theta) \propto 1$, la distribución *posterior* sería de la siguiente manera:

$$p(\theta|y) \propto \Phi(\theta)^2(1 - \Phi(\theta))^3, \quad I(\theta)_{(-3,3)} \tag{8}$$

3.2. Ejemplo B

Un escenario más general consiste en pensar que las dificultades de los ítems no son iguales y son diferentes de 0. Considere además que n estudiantes presentan una prueba con k ítems. Bajo este esquema la distribución *prior* para los parámetros de habilidad y dificultad están dadas por:

$$p(\theta, b|y) \propto p(y|\theta, b)p(\theta)p(b) \tag{9}$$

Donde:

$$p(\theta) = \prod_n N(0, \sigma^2), \quad p(b) = \prod_k N(0, \tau^2) \tag{10}$$

Con σ^2 y τ^2 hiperparámetros de las distribuciones *prior*. Siempre que se usen distribucciones *apriori* tipo gaussianas, la distribución posterior $p(\theta, b|y)$ no tiene una forma conocida, por lo cual las aproximaciones se hacen necesarias. En este caso algunos metodos como los Laplacianos resultan adecuados.

3.3. Propuesta

Para el desarrollo del presente trabajo se utilizó una distribución *prior* no informativa para el parámetro de habilidad, y una distribución prior*no informativa* para el parámetro de dificultad. Además se incluyeron hiperparámetros sobre las distribución *prior* de la del parámetro de dificultad; en otras palabras:

$$p(\theta) \sim N(0, 1), \quad p(b) \sim N(\mu_1, \tau_1^2) \tag{11}$$

Donde:

$$\mu_1 \sim N(0, 1/1000), \quad \tau_2 \sim Gamma(1/1000, 1/1000) \tag{12}$$

Cabe aclarar que las distribuciones *posterior* no tienen una forma conocida, por lo cual, en este sentido, la parte computacional es un factor clave para la comparación de los dos métodos de estimación (clásico y bayesiano).

En el caso bayesiano, el proceso de estimación de los parámetros se hizo a través del muestreador de Gibbs [2], utilizando el programa `JAGS` *(Just Another Gibbs Sampler)* [3] asociado al programa `R`. Como valores inciales, estos parámetros arrancaron en 0. Se escogió realizar 10000 simulaciones (*draws*) y se optó por una etapa de calentamiento de 1000 *draws*.

4. Resultados

En esta sección se presentan los resultados obtenidos en el proceso de calificación del componente de escucha de la aplicación 2015 de la prueba de inglés (8950 docentes licenciados en inglés) utilizando tanto el enfoque bayesiano como el clásico.

4.1. Estimación de los b y los θ

Como se realizó una sola simulación, se optó por utilizar la prueba de diagnóstico de *Densidad espectral* propuesta por Geweke & Porter-Hudak (1983), de la cual para un total de 2030 parámetros (2000 habilidades y 30 dificultades) se obtuvo tan solo una tasa de **no** convergencia del 7.68 % (figura 2).

Resulta importante señalar que, a medida que la dificultad de un ítem aumente, se espera que la habilidad de la persona que constesta dicho ítem también aumente. De esta forma, se espera que entre más personas respondan un ítem, la dificultad de este disminuya proporcionalmente. Esta relación se puede observar en la figura 3. Además, se puede observar que bajo el enfoque bayesiano se mantiene la relación entre el porcentaje de respuestas correctas de un ítem con la dificultad de este mismo, aclarando que el ítem con dificultad más pequeña es aquel que tiene un menor porcentaje de aciertos. Aunque no existe mucha diferencia, vale la pena establecer cuál de estas metodologías presenta menor error de estimación de los parámetros (ver subsección 4.2).

[2] http://halweb.uc3m.es/esp/Personal/personas/causin/esp/2012-2013/SMB/Tema8.pdf
[3] http://mcmc-jags.sourceforge.net/

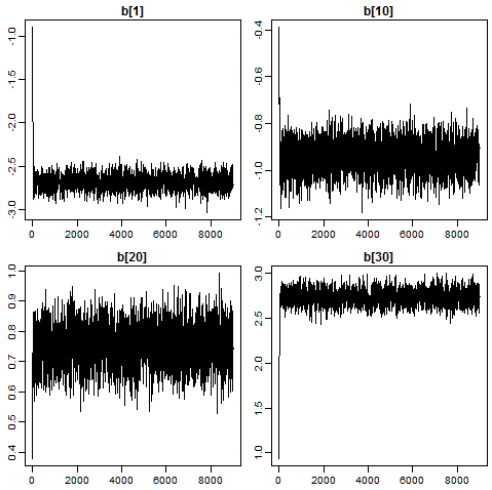

Figura 2: *Cadenas de Markov para algunos b. Fuente: elaboración propia.*

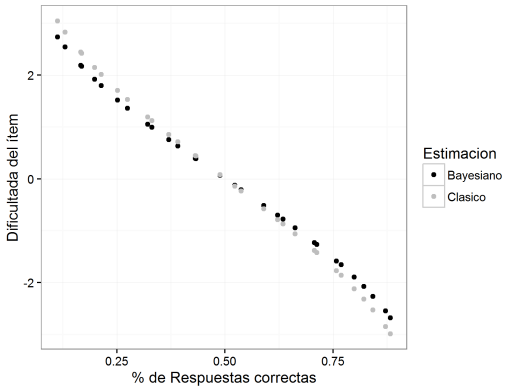

Figura 3: *Dificultad del item vs porcentaje de respuestas correctas. Fuente: elaboración propia.*

4.2. Comparación con enfoque clásico

Si el enfoque bayesiano ha de ser consistente, deberá tener al menos una buena relación con el enfoque clásico. En este sentido, la correlación entre las estimaciones deberá ser alta, aunque no necesariamente cercana a uno. En la figura 4 se puede observar que a medida que la dificultad estimada de un ítem crece en el enfoque clásico, también aumenta en el enfoque bayesiano. Nótese que el enfoque bayesiano mantiene el supuesto de que las dificultades de los ítems están centradas en 0, lo cual es muy bueno, ya que afirma que un enfoque distinto para estos modelos

clásicos funciona de manera similar.

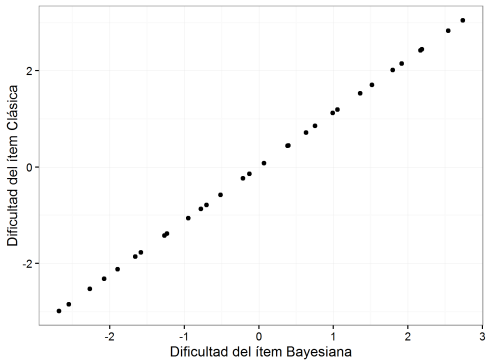

Figura 4: *Relación entre las estimaciones. Fuente: elaboración propia.*

En las figuras 5 y 6 se observa la función de densidad de las dificultades de los ítems y las habilidades de las personas. Estas funciones son muy similares, lo cual era de esperarse por lo resultados vistos anteriormente. Sin embargo, en cuanto a la habilidad, sí hay una diferencia notoria en la estimación bayesiana, sobre todo porque las habilidades se están concentrando en una mayor medida en el rango [-2,2].

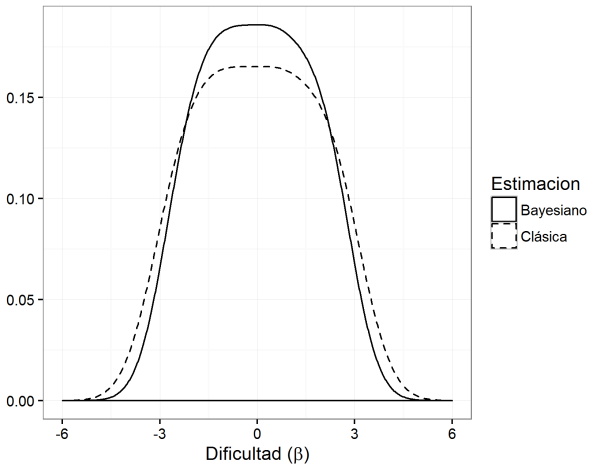

Figura 5: *Densidad de la dificultad estimada. Fuente: elaboración propia.*

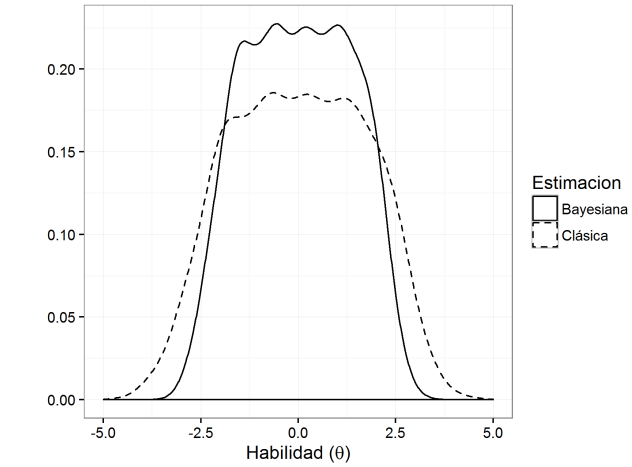

Figura 6: *Densidad de la habilidad estimada. Fuente: elaboración propia.*

Otro aspecto importante por comparar es la tendencia de las dificultades de los ítems al momento de la simulación. Cuando se simularon dichos datos, se puso como referencia que el primer ítem debería tener la dificultad más baja y el último ítem la dificultad más alta. En la figura 7 se evidencia que dicho comportamiento se mantiene. Resulta bastante particular que la estimación para los ítems 14, 15 y 16 sea la misma en ambas metodologías, mientras que a medida que la dificultad se va acercando a los extremos del rango, la diferencia en la dificultad estimada es mucho más marcada.

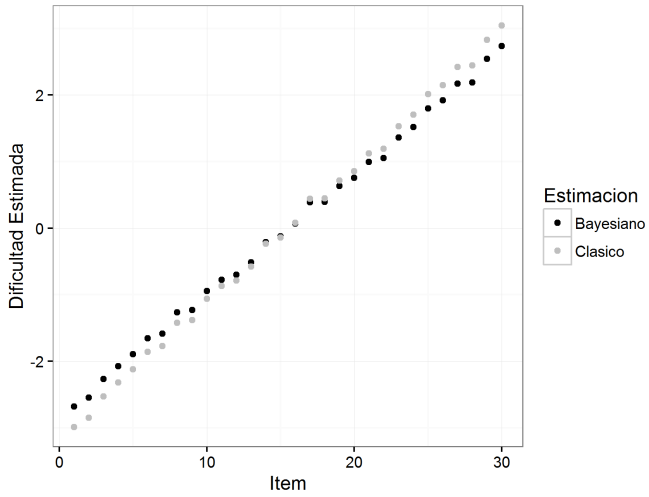

Figura 7: *Dificultad estimada del ítem. Fuente: elaboración propia.*

Más aún, el proceso de identificar cuál de las 2 metodologías produce menor error de estimación puede ser empleado como criterio para definir cuál de los dos procesos de estimación es más preciso. En las figuras 8 y 9 se observa una diferencia bastante notoria tanto para las habilidades como para las dificultades, donde el enfoque bayesiano produce un error de estimación mucho más pequeño. Cabe resaltar que la distribución del error de estimación bayesiano conserva prácticamente el mismo comportamiento en términos distribucionales que la metodología clásica.

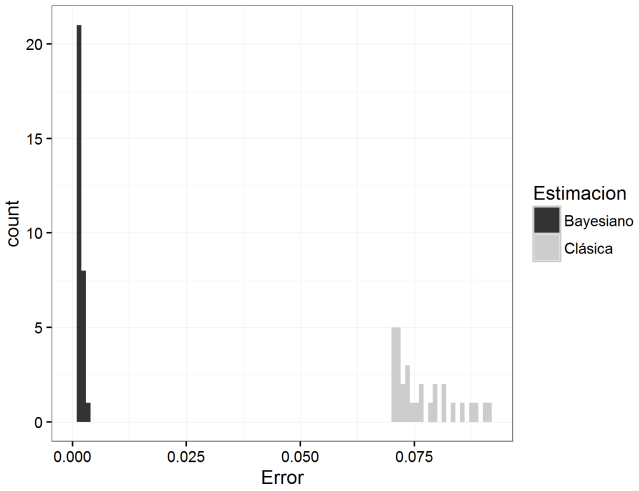

Figura 8: *Error de estimación de las dificultades. Fuente: elaboración propia.*

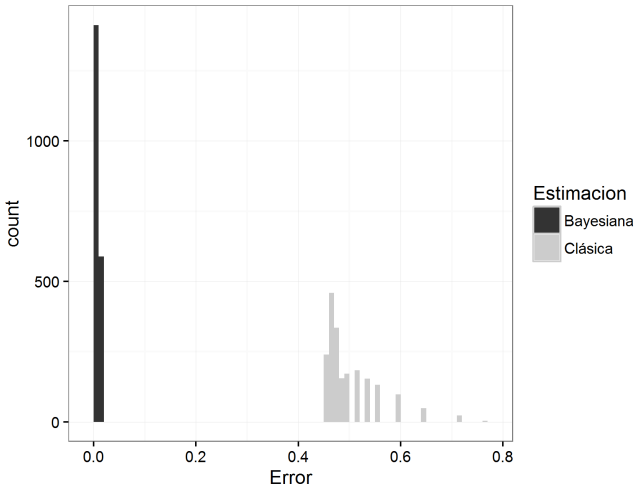

Figura 9: *Error de estimación de las habilidades. Fuente: elaboración propia.*

Tabla 1: *Estimación y error de estimación de la dificultad de los ítems. Fuente: elaboración propia.*

bayesiano			Clásico		
Estimación	Error	Item	Estimación	Error	Item
-2.6799	0.0134	1	-2.9860	0.0900	1
-2.5449	0.0160	2	-2.8460	0.0870	2
-2.2747	0.0108	3	-2.5280	0.0830	3
-2.0658	0.0104	4	-2.3210	0.0800	4
-1.8904	0.0093	5	-2.1190	0.0780	5
-1.6552	0.0079	6	-1.8600	0.0760	6
-1.5837	0.0071	7	-1.7740	0.0750	7
-1.2646	0.0063	8	-1.4250	0.0730	8
-1.2343	0.0062	9	-1.3840	0.0730	9
-0.9523	0.0060	10	-1.0630	0.0720	10
-0.7791	0.0060	11	-0.8710	0.0710	11
-0.6993	0.0048	12	-0.7900	0.0710	12
-0.5143	0.0048	13	-0.5780	0.0700	13
-0.2216	0.0040	14	-0.2360	0.0700	14
-0.1313	0.0046	15	-0.1390	0.0700	15
0.0672	0.0045	16	0.0800	0.0700	16
0.3848	0.0044	17	0.4380	0.0700	17
0.3847	0.0037	18	0.4480	0.0700	18
0.6313	0.0041	19	0.7140	0.0710	19
0.7471	0.0040	20	0.8510	0.0710	20
0.9885	0.0045	21	1.1180	0.0720	21
1.0479	0.0058	22	1.1880	0.0720	22
1.3517	0.0060	23	1.5300	0.0740	23
1.5203	0.0054	24	1.7060	0.0750	24
1.7854	0.0065	25	2.0120	0.0770	25
1.9061	0.0078	26	2.1460	0.0790	26
2.1585	0.0088	27	2.4200	0.0810	27
2.1827	0.0096	28	2.4450	0.0820	28
2.5424	0.0113	29	2.8280	0.0870	29
2.7278	0.0118	30	3.0410	0.0900	30

5. Conclusiones

La metodología bayesiana para la estimación de los parámetros de dificultad y habilidad en un modelo de *Rasch*, resultó tener un comportamiento similar al de la metodología clásica de dicho modelo, manteniendo el supuesto de que tanto las habilidades como las dificultades estan centradas en cero. Además, se encontró que la metodología bayesiana reduce el error de estimación en los parámetros del modelo de Rasch, lo que implica una ganancia en términos de precisión al momento de la calificación de los exámenes.

Otro hallazgo importante, es que aproximadamente el 92 % de las cadenas convergieron, lo cual reafirma que las estimaciones son factibles. Por último, a pesar de que esta metodología demanda una mayor cantidad de tiempo y capacidad computacional, la ventaja en términos de la reducción del error es suficiente para pensar en la implementación de esta metodología a los procesos clásicos de calificación que actualmente realiza el Icfes.

Recibido: 5 de marzo del 2015
Aceptado: 18 de abril del 2015

Referencias

Ayala, R. (2008), *The theory and pratice of item response theory*, 1 edn, The Guilford Press.

Fox, J. (2010), *Bayesian Item Response Modeling. Theory and Applications*, 1 edn, Springer.

Geweke, J. & Porter-Hudak, S. (1983), 'The estimation and application of long-memory times series models'.

Mislevy, R. (1986), 'Bayes model estimation in item response models', *Psychometrika* **51**(1), 177–195.

Rasch, G. (1980), *Probabilistic models for some intelligence and attainment tests*, 1 edn, University of Chicago Press.

Rigdon, S. & Tsutakawa, R. (1983), 'Parameter estimation in latent trait models', *Psychometrika* **48**(1), 567–574.

Sinharay, S. (2003), Bayesian item fit analysis for dichotomous item response theory models, Technical report, ETS, Princeton, NJ 08541.

Swaminathan, H. & Gifford, J. A. (1982), 'Bayesian estimation in the rasch model', *Journal of Educational Statistics* **7**(1), 175–192.

A. Códigos

```
rm(list = ls())
library(boot)
library(xtable)
library(ggplot2)
library(reshape2)
library(dplyr)
library(data.table)

p <- 30 # Numero de \'items
n <- 2000 # Numero de estudiantes

# Habilidad para cada personas
theta = seq(from = -3, to = 3, length.out = n)
# dificultad para cada \'item
b <- seq(from = -3, to = 3, length.out = p)
# Matriz para crear los 1 y 0
pr <- y <- matrix(NA, nrow = n, ncol = p)

set.seed(11102015)
# Construccion de las respuestas deacuerdo a un modelo de Rasch
for(i in 1:p){
x <- theta - b[i]
pr[, i] <- inv.logit(x)
y[, i] <- rbinom(n, 1, pr[, i])
}

Rasch.data <- matrix(y, ncol = length(b))

PropNRC <- data.frame(Item = seq(from = 1, to = p, by = 1))
PropNRC[,"Proporcion_de_respuestas_correctas"] <- apply(Rasch.data,
2, mean)

# xtable(PropNRC, caption = "% de respuestas correctas por Item",
digits = 3)
GraphPropNRC <- ggplot() + geom_bar(data = PropNRC,
aes(y = Proporcion_de_respuestas_correctas, x = Item),
stat = "identity", fill = "darkblue") +
ylab("% Respuestas Correctas")
# ggsave(plot = GraphPropNRC, filename = "../GraphPropNRC.png")

# NRC de los estudiantes
PersoNRC <- data.frame(Persona = seq(from = 1, to = n, by = 1))
PersoNRC[,"NRC"] <- as.numeric(apply(Rasch.data, 1, sum))
#summary(PersoNRC[,"NRC"])
```

```
PropPerNRC <- prop.table(table(PersoNRC[,"NRC"]))*100
# xtable(PropPerNRC,
#  caption = "% de estudiantes deacuerdo al n\'umero de
respuestas correctas")
PropPerNRC <- data.frame(PropPerNRC)
colnames(PropPerNRC) <- c("NRC","Estudiantes")

GraphEstuNRC <- ggplot() + geom_bar(data = PropPerNRC,
aes(y = Estudiantes, x = NRC),
stat = "identity", fill = "darkblue") + ylab("% de Estudiantes")
# ggsave(plot = GraphEstuNRC, filename = "../GraphEstuNRC.png")


############################################################
# # Estimacion Cl\'asica
############################################################
library(mirt)

raschfit  <- mirt(Rasch.data, model = 1, itemtype='Rasch', SE = TRUE)

#########################
# dificultad de los \'items
#########################

dificult  <- coef(raschfit, CI = 0.99, printSE = TRUE,
digits = 3, as.data.frame = TRUE)

filtroUno <- substr( rownames(dificult),8,8)
filtroDos <- substr( rownames(dificult),9,9)
dificult  <- subset(dificult, filtroUno == 'd' | filtroDos == 'd')
dificult  <- as.data.frame(dificult)
dificult[,"par"] <- dificult[,"par"]*-1 #se multiplica por negativo,
# ya que la salida de mirt es con signo contrario

Item <- seq(from =1, to = ncol(Rasch.data), by = 1)
Dificultad <- cbind(Item = Item, dificult)
names(Dificultad) <- c("Item","Dificultad","Error")
row.names(Dificultad) <- NULL
xtable(Dificultad)

# Habilidad de las personas
scores  <- fscores(raschfit, method = 'EAP', full.scores=TRUE,
full.scores.SE = TRUE) # Habilidades
scores <- data.frame(scores)
names(scores) <- c("Habilidad", "Error")
row.names(scores) <- NULL
```

```
xtable(scores[1:10,])

######################
# # %NRC vs dificultad
######################
NRC <- apply(Rasch.data, 2, sum)
PorNRC <- NRC/n
NRCDif <- cbind(Dificultad,PorNRC)
NRCDifGra <- ggplot(data = NRCDif) + geom_point(aes(x= Dificultad,
y = PorNRC)) +  ylab("% NRC")
# ggsave(plot = NRCDifGra, filename = "../NRCDifGra.png")

############################################################
# # Simulacion bayesiana del modelo de Rasch
############################################################

# codigo de WinBugs

library(R2jags)
library(coda)
library(lattice) # graficas
library(R2WinBUGS)
library(superdiag) # criterio de convergencia de MCMC
library(mcmcplots) # graficas

# Definicion de elementos

Y <- Rasch.data # Matriz de 1's y 0's
burnin <- 100 # Burning
iter   <- 1000 # Simulaciones
chain  <- 1 # Cadenas
thin   <- 1 # Saltos para la seleccion

Rasch.model <-function() {
for (i in 1:n){
for (j in 1:p){
Y[i, j] ~ dbern(prob[i, j])
logit(prob[i, j]) <-  ( theta[i] - b[j])
}
theta[ i ] ~ dnorm(0, 1)
}
for(j in 1:p){
b[j] ~ dnorm(mu[j], tau[j])
mu[j] ~ dnorm(0, 1/10000)
tau[j] ~ dgamma(1/10000, 1/10000)
}
```

```
}

Rasch.data  <- list("Y", "n", "p")
Rasch.param <- c("theta", "b")
Rasch.inits <- function(){list("theta"=rep(0,n), "b"=rep(0,p))}
set.seed(123)

Rasch.fit <- jags(data = Rasch.data, inits = Rasch.inits,
Rasch.param, n.chains = chain, n.iter = iter,
n.burnin = burnin, n.thin = thin,model.file = Rasch.model)

############################################################
# # Criterios de convergencia
############################################################
bayes.mod.fit.mcmc <- as.mcmc(Rasch.fit)
geweke <- geweke.diag(bayes.mod.fit.mcmc)
geweke <- data.frame(Z_Value = unlist(geweke),
 param = names(unlist(geweke)))
noconver <- which(geweke[,"Z_Value"] > qnorm(0.975) |
geweke[,"Z_Value"] < qnorm(0.025) )
NoConvergencia <- geweke[noconver,]
TasaNoConver <- nrow(NoConvergencia) / (nrow(geweke)-3)
TasaNoConver
# Gelman_Rubin <- gelman.diag(bayes.mod.fit.mcmc, transform=TRUE)
png("../Chain_betas.png")
traplot(bayes.mod.fit.mcmc,parms = c("b[1]","b[10]","b[20]","b[30]"))
dev.off()
d <- summary(bayes.mod.fit.mcmc)
# superdiag(mcmcoutput = bayes.mod.fit.mcmc, burnin = 100 )

############################################################
# # Extraccion de los parametros a estimar
############################################################
resumen <- as.data.frame(d$statistics)
resumen[,"Parametro"] <- row.names(resumen)
resumen[,"Parametro"] <- gsub("\\[|\\]","", resumen[,"Parametro"])
betasEst    <- subset(resumen, Parametro %like% "b")
thetasEst   <- subset(resumen, Parametro %like% "theta")
devianceEst <- subset(resumen, Parametro %like% "deviance")
############################################################
# # Coherencias de las estimaciones
############################################################
propNRC <- apply(Y, 2, mean)
verifica <- data.frame(PropNRC = propNRC,
 bayesiano = betasEst[,"Mean"],
 Clasico = dificult[,"par"])
```

```
colnames(verifica) <- c("PropNRC", "bayesiano", "Clasico")

ggplot() + geom_point(data = verifica, aes(x=bayesiano, y = Clasico),
colour ="green") +
xlab("Dificultad del \'item bayesiana") +
ylab("Dificultad del \'item Cl\'asica")
ggsave(plot = last_plot(), filename = "../Verificacion_1.png")

verifica <- melt(verifica, id.vars = "PropNRC")
names(verifica) <- c("PropNRC","Estimacion", "Valor")
ggplot() + geom_point(data = verifica,
aes(x=PropNRC, y = Valor, colour = Estimacion)) +
 xlab("% de Respuestas correctas") +
 ylab("Dificultada del \'item")
ggsave(plot = last_plot(), filename = "../Verificacion_2.png")

############################################################
# # Comparacion de las metodolog\'ias
############################################################
names(scores) <- c("Mean","Error")
scores[,"Persona"] <- seq(from = 1, to = nrow(scores), by = 1)
scores[,"Estimacion"] <- "Cl\'asica"
thetasEst <- thetasEst[,c("Mean","Time-series SE","Parametro")]
row.names(thetasEst) <- NULL
colnames(thetasEst) <-c("Mean","Error","Persona")
thetasEst[,"Estimacion"] <- "bayesiana"
thetas <- rbind(scores, thetasEst)
thetas[,"Persona"] <- gsub("theta","",thetas[,"Persona"])
row.names(thetas) <- NULL

# # Dificultad
dificult[,"Parametro"] <- seq(from = 1, to = nrow(dificult), by = 1)
colnames(dificult) <- c("Mean", "Error", "Item")
betasEst <- betasEst[,c("Mean","Time-series SE","Parametro")]
colnames(betasEst) <- c("Mean", "Error", "Item")
betasEst[,"Estimacion"] <- "bayesiano"
dificult[,"Item"] <- gsub("b", "",dificult[,"Item"])
dificult[,"Estimacion"] <- "Cl\'asica"
betas <- rbind(betasEst, dificult)

# Densidades de la habilidad
ggplot() + geom_density(data = thetas, aes(x =  Mean,
 colour = Estimacion ))+
xlab(expression(paste("Habilidad (",theta,")"))) + ylab("")+
xlim(-5,5)
```

```
ggsave(plot = last_plot(), filename = "../Densidad_Habilidad.jpg")

# Densidades de la dificultad
ggplot() + geom_density(data = betas, aes(x =  Mean,
 colour = Estimacion ))+
 xlab(expression(paste("Dificultad (",beta,")"))) + ylab("")+
 xlim(-6,6)

ggsave(plot = last_plot(), filename = "../Densidad_Dificultad.jpg")

# Error de la estimaciones habilidades
thetas %>%
ggplot() + geom_histogram(aes(x=Error, fill = Estimacion ))
ggsave(plot = last_plot(), filename = "../Error_Habilidades.png")

# Error de la estimaciones dificultades
betas %>%
ggplot() + geom_histogram(aes(x=Error, fill = Estimacion ))
ggsave(plot = last_plot(), filename = "../Error_Dificultades.png")

# Error de la estimaciones dificultades
betas[,"Item"] <- gsub("b","",betas[,"Item"])
betas %>%
mutate(Item = as.numeric(Item)) %>%
arrange(Item) %>%
ggplot() + geom_point(aes(x=Item, y=Mean, colour = Estimacion )) +
ylab("Dificultad Estimada")
ggsave(plot = last_plot(), filename = "../Dificultades_Item.png")

xtable(cbind(betas[1:30,-4],betas[31:60,-4]),digits = 4)
```