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. 55–78

Una aplicación de valores plausibles a la calificación de pruebas estandarizadas vía simulación[1]

An aplication of plausible values to the standaridzed test scoring through simulation

Michel Felipe Córdoba Perozo[a]
mcordoba@contratista.icfes.gov.co

Resumen

El uso de los valores plausibles en evaluaciones estandarizadas de gran escala desempeña el papel de imputación que se requiere cuando el marco de referencia es muy grande y cada individuo no aborda la totalidad de los ítems puestos en producción. En dicho caso se recurre a la definición de un diseño por bloques que garantice una cuota adecuada de individuos evaluados por cada ítem para que el marco de referencia sea abordado con suficiencia por toda la población objeto de estudio. En general, el método de imputación consiste en encontrar la distribución *a posteriori* del rasgo latente que es asociado a la habilidad del individuo, mediante la ponderación de la distribución que induce el modelo de teoría de respuesta al ítem y una regresión latente asociada a algunas variables medidas en el individuo. El presente artículo muestra un ejemplo de simulación, donde se pueden observar de manera sencilla las bondades que brinda el método en los resultados agregados bajo este esquema de aplicación particular.

***Palabras clave*:** valores plausibles, imputación de datos, rasgo latente, pruebas estandarizadas .

Abstract

The use of Plausible Values in large scale standarized tests, develop the role of imputation when the reference framework is too large and each person can not adress the totality of the items in the production. In that case appleal to block design definition that ensures an appropiate share of individuals per item aiming

[1]Córdoba, M. F. (2016) Una aplicación de valores plausibles a la calificación de pruebas estandarizadas vía simulación. *Comunicaciones en Estadística*, **9**(1), 55-78.

[a]Msc. Estadística. Subdirección de Estadística, Icfes, Colombia

that the framework is addressed smugly across the study population is necessary. In general the imputation methos consist in find the posterior distribution of the feature latent that is associated to the individual hability, by weighting distribution that induces model item response theory and regression associated with some latent variables measured in the individual. This paper shows an example of simulation where you can easily see the advantages offered by the method in the aggregate results of this particle scheme application

***Keywords*:** generalized linear models, rating equalization, Beta regression models, standardized test .

1. Introducción

En las evaluaciones de gran escala, el marco de referencia que define una prueba puede ser demasiado amplio. Esto implica que el número de preguntas que debe abordar el total de la población es tan grande que dificulta que cada individuo aborde todas y cada una de las mismas. Esta dificultad estimula el uso de un diseño que asigne de manera aleatoria a cada individuo solamente una muestra de los ítems de forma tal que uno de ellos sea abordado por un número adecuado de individuos.

Como cada pregunta no es abordada por todos los individuos de la población, el diseño induce un error producto de las observaciones perdidas o ausentes en la estimación de las estadísticas que definen la habilidad promedio de la población. Este problema puede abordarse mediante la implementación de alguna técnica de imputación, no sobre las respuestas de los ítems sino sobre la habilidad de los individuos para propósitos de agregación. La técnica más reconocida es la de los valores plausibles, el cual usa el método de imputación múltiple.

En el presente artículo se ilustra, mediante un pequeño ejercicio de simulación, el uso y la utilidad de esta técnica: en el segundo apartado se describe en qué consiste la teoría que abarca desde el ajuste de los parámetros del modelo de respuesta al ítem hasta la imputación múltiple que genera los valores plausible. En el tercero apartado se simula una población de individuos en donde se calculan los resultados agregados en tres escenarios distintos para visualizar el aumento del sesgo y la disminución de la variablidad observada cuando cada individuo deja de presentar una parte de la prueba. Por último, se relaciona la solución de este problema cuando se usan los valores plausibles.

2. Marco teórico

2.1. Modelos usuales unidimensionales de teoría de respuesta al ítem

Los modelos de teoría de repuesta al ítem (TRI) son un caso especial de los modelos lineales generalizados (MLG). Estos establecen una relación entre las respuestas a un conjunto de ítems de un individuo a quien se le aplica una prueba y el rasgo latente de este, medido sobre alguna escala definida, conocido por muchos autores como habilidad. De manera específica, la probabilidad de que cierto individuo acierte a un ítem se asume como una función de θ, el símbolo usado para denotar la característica que se quiere medir. En esta sección se describen las características más generales de los modelos que usualmente se pueden encontrar en la práctica (Hulin et al. 1983).

2.1.1. El modelo de Rasch o logístico de un parámetro (1PL)

El modelo de Rasch es el modelo más simple de la familia de modelos unidimensionales. Este modelo usa solamente un parámetro para caracterizar cada ítem y un parámetro para caracterizar cada persona. Para el individuo i con habilidad latente θ_i asumida unidimensional, la probabilidad de obtener una respuesta acertada al item j, es decir, $P(Y_{ij} = 1|\theta_i)$ con Y_{ij} una variable binaria que toma el valor 1 si el individuo i responde correctamente el ítem j, es dada por la siguiente ecuación:

$$P(Y_{ij} = 1|\theta_i) = P_j(\theta_i) = \frac{1}{1 + \exp[-K(\theta_i - b_j)]} \tag{1}$$

En donde b_j es el parámetro que define la dificultad del ítem: cuando b_j incrementa, la probabilidad de contestar correctamente el ítem j decrece y K es una constate que define la escala, usualmente definida como 1.702. Es importante señalar que cuando θ_i es realmente unidimensional, el rasgo latente no es más que un escalamiento no lineal de la estadística τ: el puntaje obtenido por teoría clásica.

Una característica de este modelo es que la probabilidad de responder correctamente cualquier ítem tiende a cero para pequeños valores de θ_i. Esto implica que los individuos con valores pequeños de θ no tendrán oportunidad de responder de manera correcta un ítem con alta o moderada dificultad, es decir, según este modelo, estos individuos no tendrán la posibilidad de acertar por azar con la respuesta correcta. Otra característica importante de este modelo es que si la dificultad fuera la misma para todos, el cambio en la probabilidad de responder de forma acertada en función de θ sería la misma. Esto implica que todos los ítems considerados tienen la misma discriminación (Hulin et al. 1983).

2.1.2. El modelo logístico de dos parámetros (2PL)

Este modelo adiciona al dartículo anteriormente un segundo parámetro que permite más flexibilidad en el ajuste de la probabilidad de responder correctamente un conjunto de ítems cuando estos tienen diferente discriminación. La expresión está dada de la siguiente manera:

$$P(Y_{ij} = 1|\theta_i) = P_j(\theta_i) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} \tag{2}$$

Donde b_j modela la dificultad del ítem j como en el caso anterior y a_j la discriminación. Este último parámetro controla la tasa de cambio de la probabilidad de responder el ítem de manera acertada en una vecindad del espacio de θ alrededor del valor b_j. Al igual que el modelo dartículo anteriormente, en este caso la probabilidad de responder correctamente cualquier ítem tiende a cero para pequeños valores de θ_i, es decir, la probabilidad de responder correctamente cualquier ítem debido al azar tiende a cero (Hulin et al. 1983).

2.1.3. El modelo logístico de tres parámetros (3PL)

Además de los dos parámetros por ítem que definen al modelo 2PL, el modelo 3PL incorpora un tercer parámetro en la ecuación, denominado el parámetro de azar o acierto casual. Su forma funcional está dada por la siguiente expresión:

$$P(Y_{ij} = 1|\theta_i) = P_j(\theta_i) = c_i + \frac{1 - c_i}{1 + \exp[-a_j(\theta_i - b_j)]} \tag{3}$$

El parámetro c_i permite adicionar flexibilidad en el ajuste de los datos, ya que modela la probabilidad de que un individuo con baja habilidad responda de manera correcta el ítem j. Si un individuo tiene habilidad baja, la probabilidad de que responda correctamente un ítem de dificultad moderada o alta es baja. En caso de que dicho individuo responda de manera correcta un ítem de estas características, el evento puede no ser debido a su habilidad sino atribuible al azar.

El modelo de tres parámetros es uno de los casos más generales de los modelos de TRI unidimensionales: en particular generaliza los modelos 1PL y 2PL. A partir de la expresión (3), para obtener un modelo 2PL basta con hacer $c_j = 0$ y hacer $a_i = K$ para caer en el caso de un modelo 1PL.

2.1.4. Otros modelos unidimensionales

Los modelos dartículos anteriormente se definen como dicotómicos puesto que la medición de la respuesta solo tiene dos posibilidades: acierto o fracaso. Por su estructura, algunos ítems pueden ser ajustados usando un modelo que considere que la elección de una opción de respuesta incorrecta pueda dar crédito parcial a la tarea que considera medir el ítem. Un caso particular de ellos es conocido como el

modelo de crédito parcial (Aitkin & Aitkin 2011). Incluyendo este, existe además una serie de modelos que relacionan la habilidad del individuo y la probabilidad de responder cada una de las múltiples opciones. Entre ellos está el modelo de respuesta graduada o el modelo de respuesta nominal. Esta clase de modelos se definen como politómicos.

Por propósitos del presente artículo, solamente se usará el modelo 2PL, como recurso para la calificación de los resultados de la prueba Saber 359 que se aplica en Colombia de manera estandarizada y periódica desde 2009 para los grados quinto y noveno y para el grado tercero a partir del 2012. Para una revisión exhaustiva de todos los modelos citados aquí, el lector puede referirse a Aitkin & Aitkin (2011) y Hulin et al. (1983).

2.1.5. Estimación de los parámetros del modelo

La estimación de los parámetros que definen el modelo es realizada vía máxima verosimilitud marginal (Bock & Aitkin 1981, Harwell et al. 1988). En general, el método asume independencia condicional en las respuestas a diferentes ítem para individuos de la misma habilidad (Kass & Steffey 1989). Sea $\mathbf{Y}_i$ el patrón de respuestas para los n ítems de un individuo i con habilidad θ_i:

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$$

Donde Y_{ij} toma el valor de 1 si el evaluado i responde de manera correcta la pregunta j y cero cuando no. La probabilidad de que el individuo responda el patrón en mención dada su habilidad θ_i es expresada como sigue:

$$P(\mathbf{Y}_i|\theta_i) = \prod_{j=1}^{n} [P(Y_{ij} = 1|\theta_i)]^{Y_{ij}} [P(Y_{ij} = 0|\theta_i)]^{1-Y_{ij}}. \tag{4}$$

Sea $g(\theta)$ la función de densidad que determina la distribución de la habilidad de todos los individuos en la población bajo estudio. La probabilidad marginal de que cualquier individuo de la población obtenga un patrón de respuestas $\mathbf{Y}$ está dada por:

$$P(\mathbf{Y}) = \int_{\mathbb{R}} P(\mathbf{Y}|\theta) g(\theta) d\theta \tag{5}$$

Esta integral no puede ser tratada analíticamente, pero la probabilidad marginal puede ser aproximada mediante la cuadratura gaussiana:

$$\bar{P}_{\mathbf{Y}} \approx \sum_{k=1}^{q} P(\mathbf{Y}|\mathbf{Y}_k) A(\mathbf{Y}_k) \tag{6}$$

Donde q es el número de puntos en la cuadratura, $\mathbf{Y}_k$ es un punto de cuadratura y $A(\mathbf{Y}_k)$ es un peso positivo correspondiente a la función de densidad $g(\cdot)$. En el

método de estimación dartículo, los valores de los ítemsse escogen de forma tal que maximicen el logaritmo de la función de verosimilitud marginal definido por:

$$\log L_M = \sum_{l=1}^{S} r_l \log \bar{P}_{\mathbf{Y}_l} \tag{7}$$

Donde r_l es la frecuencia con la cual el patrón $\mathbf{Y}_l$ es observado en una muestra de N estudiantes y S es el número de patrones distintos observados. El algoritmo EM junto con los métodos Newton-Gauss o Fisher-Scoring son usados para resolver las ecuaciones implícitas necesarias para encontrar la solución de la derivada $\partial \log L_M / \partial \Pi_j$ con $\Pi_j = (a_j, b_j)$

2.1.6. Estimación de la distribución latente de θ

Para lograr la estimación de la distribución latente, es necesario encontrar sus medidas indeterminadas de localización y escala. Esta indeterminación surge porque en el logit

$$z_j = a_j(\theta - b_j) \tag{8}$$

cualquier cambio en el origen de θ puede ser controlado por b_j y cualquier cambio en la unidad de θ puede ser controlado por a_j. Para establecer el parámetro de localización, se fija la media de la distribución latente en 0 y para establecer el parámetro de escala se fija la varianza de la distribución latente en 1.

Una forma conveniente de caracterizar una distribución latente arbitraria con media y varianza finitas es calcular la densidad de probabilidad en un número finito de adecuadas elecciones de valores de θ y normalizar las densidades dividiendo por el total. Esos valores normalizados pueden ser usados como los pesos $A(\mathbf{Y}_k)$ de la cuadratura (6).

2.1.7. Puntaje o habilidad del individuo (*test score*)

En esta sección se presentan los enfoques clásico vía máxima verosimilitud y el enfoque bayesiano para ajustar la habilidad de los individuos.

Estimación vía máxima verosimilitud

La estimación vía máxima verosimilitud del rasgo latente del i-ésimo estudiante, denotada por $\hat{\theta}_i$, es el valor de θ_i que maximiza la siguiente función:

$$\log L(\theta_i) = \sum_{j=1}^{n} \{ y_{ij} \log P(Y_{ij} = 1|\theta_i) + (1 - y_{ij}) \log P(Y_{ij} = 0|\theta_i) \}. \tag{9}$$

La ecuación que se debe resolver es:

$$\frac{\partial \log L(\theta_i)}{\partial \theta_i} = \sum_{j=1}^{n} \frac{y_{ij} - P(Y_{ij} = 1|\theta_i)}{[P(Y_{ij} = 1|\theta_i)][P(Y_{ij} = 0|\theta_i)]} \frac{\partial P(Y_{ij} = 1|\theta_i)}{\partial \theta_i} = 0 \tag{10}$$

El estimador de máxima verosimilitud $\hat{\theta}_i$ es calculado por Fisher Scoring, que depende de la información de Fisher que se obtiene así:

$$I(\theta_i) = \sum_{j=1}^{n} a_j^2 P(Y_{ij} = 1|\theta_i)[P(Y_{ij} = 0|\theta_i)]. \tag{11}$$

Escrita de forma general, la solución de las iteraciones del método es esta:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + I^{-1}(\hat{\theta}_t) \left(\frac{\partial \log L(\theta)}{\partial \theta} \bigg|_{\theta=\hat{\theta}_t} \right). \tag{12}$$

El error estándar de las estimaciones vía máxima verosimilitud es dado por:

$$S.E.(\hat{\theta}) = \sqrt{1/I(\hat{\theta})}. \tag{13}$$

Una desventaja de este método es que no hay solución a las ecuaciones de estimación cuando el estudiante responde de manera correcta o de manera incorrecta absolutamente todos los itemas. Estos problemas no surgen en los otros métodos de estimación.

Estimación Bayesiana

El estimador Bayesiano es el promedio de la distribución *a posteriori* de θ dado que se ha observado el patrón de respuestas $\mathbf{Y}$. Este puede ser aproximado por cuadratura gaussiana de la siguiente forma:

$$\hat{\theta}_i \cong \frac{\sum_{k=1}^{q} \mathbf{Y}_k P(\mathbf{Y}|Y_k) A(\mathbf{Y}_k)}{\sum_{k=1}^{q} P(\mathbf{Y}|Y_k) A(\mathbf{Y}_k)} \tag{14}$$

Este estadístico es llamado el estimador esperado *a posteriori*. Una medida de su precisión es la desviación estándar *a posteriori* (DEP) aproximada mediante la siguiente expresión:

$$DEP(\hat{\theta}_i) \cong \frac{\sum_{k=1}^{q} (\mathbf{Y}_k - \hat{\theta}_i)^2 P(\mathbf{Y}|\mathbf{Y}_k) A(\mathbf{Y}_k)}{\sum_{k=1}^{q} P(\mathbf{Y}|\mathbf{Y}_k) A(\mathbf{Y}_k)} \tag{15}$$

Los pesos $A(\mathbf{Y}_k)$ dependen de la distribución asumida para θ. Hay posibilidades de tener pesos teóricos o pesos empíricos $A^*(\mathbf{Y}_k)$ o pesos subjetivos.

Este estimador existe para cualquier patrón de respuestas y tiene un error promedio más bajo en la población que cualquier otro estimador, incluso que el estimador de máxima verosimilitud. En general, es sesgado hacia la media de la población pero el sesgo es pequeño (Bock & Mislevy 1982).

El método usado en la prueba Saber 359 es el bayesiano. La distribución *a priori* del parámetro θ es considerada normal y el número de puntos de la cuadratura definido es 30.

2.2. Evaluación de la educación a gran escala

Una evaluación a gran escala se encarga de medir lo que los integrantes de una población específica saben y pueden hacer con respecto a ciertas competencias relacionadas con algunos tópicos de interés de algún programa académico o si han adquirido habilidades necesarias para realizar actividades en el futuro. La amplitud de los temas medidos en estos programas es tal que un número muy grande de contenidos y habilidades se evalúan (González & Rutkowski 2010). Hay muchas evaluaciones de este tipo, pero como ejemplo de estas pruebas se puede mencionar la prueba Saber 359 en Colombia que mide a los estudiantes de todo el país sus competencias en lenguaje, matemáticas, ciencias naturales y competencias ciudadanas en los grados de tercero, quinto y noveno, o la prueba PISA (Programme for International Student Assessment), que evalúa cada tres años las competencias en matemáticas, en lectura y en ciencias naturales a niños de quince años.

Las pruebas estandarizadas a gran escala, por lo general, se proponen evaluar un extenso dominio de contenido académico en la población. Debido a esto, para evitar sobrecargar a los estudiantes y por costos económicos y de tiempo, las pruebas están diseñadas de tal forma que a un estudiante se le administra solamente una fracción de esta, es decir, una combinación particular de ítems de la prueba de forma tal que asegure la cobertura necesaria en toda la población. Este tipo de estructura se conoce como diseño de muestreo matricial múltiple de ítems (*multile matrix sampling*) o simplemente diseño de muestreo matricial de ítem. Este enfoque permite estimar de forma precisa el comportamiento de las competencias en la población o en subpoblaciones y, a su vez, permite la cobertura de todo el marco de referencia de la evaluación. Por otro lado, permite reducir la carga de cada estudiante evaluado y el tiempo que dura la prueba en su ejecución (González & Rutkowski 2010).

Dado que cada estudiante no presenta toda la prueba sino solamente una parte de ella, la precisión de la medición individual es sacrificada por el interés de tener la capacidad de evaluar todo el marco de referencia en la población. Este énfasis hace que este diseño no sea óptimo para reportar resultados individuales, por lo que los resultados que se obtienen en este tipo de evaluaciones siempre es agregado. Por ejemplo, en Saber 359, el mínimo nivel de agrupación al que los resultados están reportados es el de establecimiento educativo. El máximo, por supuesto, es el nivel de agregación nacional. Por conveniencia operativa, los ítems que componen el instrumento de medición son asignados a bloques que son luego combinados en formas de acuerdo con una especificación particular o al diseño. Para una revisión exhaustiva del diseño en bloques, el lector puede referirse a (González & Rutkowski 2010).

En el caso particular de grado quinto en la prueba Saber 359 del 2015 se evaluaron las áreas de ciencias naturales y lenguaje y el área de matemáticas. Para completar la medición del marco de referencia se construyeron 6 bloques. Cada bloque consta de 24 ítems, es decir, se consolidaron 144 ítems para la evaluación. Por diseño, a cada estudiante que presentó la prueba de matemáticas se suministró solamente

dos bloques de seis. Para una revisión exhaustiva del diseño de la prueba Saber 359, el lector puede referirse al informe técnico de la prueba (*Informe técnico SABER 5o. y 9o. 2009* n.d.).

2.3. Valores plausibles - imputación múltiple

La presente sección muestra de manera general en qué consiste el método de valores plausibles, el modelo de calificación que induce y su caracterización en términos matemáticos. El método de los valores plausibles imputa sobre los resultados agregados esa información que por diseño no puede ser medida en toda la población.

2.3.1. Modelo de calificación

Suponga ahora que además se dispone de información auxiliar $\mathbf{X}$ medible en cada uno de los individuos de la población. Generalmente, el objetivo se centra en encontrar la distribución asociada con la habilidad del estudiante de la siguiente manera:

$$P(\theta|\mathbf{X},\mathbf{Y}) \tag{16}$$

En este caso, la distribución de la habilidad en la población no solamente está condicionada sobre el patrón de respuestas $\mathbf{Y}$, sino que ahora tiene asociación con los valores de las variables exógenas del individuo $\mathbf{X}$. La expresión (16) puede escribirse de la siguiente forma:

$$P(\theta|\mathbf{X},\mathbf{Y}) = P(\mathbf{Y}|\mathbf{X},\theta)P(\theta|\mathbf{X}). \tag{17}$$

Para esto es necesario hacer los siguientes supuestos, conocidos como los supuestos de la independencia condicional:

1. $\mathbf{Y}$ es condicionalmente independiente de $\mathbf{X}$, es decir:

$$P(\mathbf{Y}|\mathbf{X},\theta) = P(\mathbf{Y}|\theta) \tag{18}$$

2. Los elementos del vector $\mathbf{Y}$ son condicionalmente independientes, es decir:

$$P(\mathbf{Y}|\theta) = \prod_{i=i}^{n} P(Y_i = y_i|\theta) \tag{19}$$

   En el caso de un modelo de dos parámetros 2PL, esta expresión no es otra cosa que el producto de probabilidades expresadas en su forma funcional como en (3) para y_i tomando valores de 0 y 1 .

Del primer supuesto y por teorema de Bayes se obiente la siguiente expresión para la ecuación (17):

$$P(\theta|\mathbf{X},\mathbf{Y}) = P(\mathbf{Y}|\theta)P(\theta|\mathbf{X}) \tag{20}$$

2.3.2. Modelo del rasgo latente

De la ecuación (20) se encuentra que falta determinar un modelo adecuado para la expresión $P(\theta|\mathbf{X})$. Para este se asume una distribución normal con matriz de varianzas Σ y media dada por una función lineal de $\mathbf{X}$, es decir:

$$P(\Theta|\mathbf{X}) = \Phi(\Theta; \mathbf{X}\Gamma, \Sigma) \tag{21}$$

Con Θ el vector de habilidades de toda la población y $\Phi(\cdot)$ la función de densidad de probabilidad de una variable aleatoria con distribución normal. Esto sugiere el siguiente modelo de regresión:

$$\Theta = \mathbf{X}\Gamma + \epsilon \tag{22}$$

Donde $\epsilon \sim N(\mathbf{0}, \Sigma)$ y Γ y Σ son parámetros a estimar. La cantidad no observable θ para cada individuo de la población depende en alguna medida de ciertas variables $\mathbf{X}$ que caracterizan la población.

Estimación de los parámetros del modelo - algoritmo EM

Como la expresión (22) no es un modelo de regresión ordinario, no se puede aplicar cualquier método para encontrar las estimaciones de los parámetros Γ y Σ. Para esto se pueden ver los valores plausibles como mediciones no observadas en un individuo. En ese escenario, para encontrar estimaciones de máxima verosimilitud se debe usar el algoritmo EM.

El primer paso del algoritmo es formular las soluciones si el vector Θ o vector respuesta es observado. Sea $\mathbf{D}$ la matriz diagonal que guarda en sus entradas los pesos de muestreo de los N individuos que induce un diseño de muestreo $\mathbf{D} = \text{diag}(w_1, \ldots, w_N)$ y sea p el número de covariables observadas. Los estimadores de máxima verosimilitud de los parámetros son:

$$\hat{\Gamma} = (\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}\Theta \tag{23}$$

y

$$\hat{\Sigma} = \frac{1}{\text{tr}(\mathbf{D})}(\Theta - \mathbf{X}\Gamma)^T\mathbf{D}(\Theta - \mathbf{X}\Gamma) \tag{24}$$

Esto permite establecer que en el k-ésimo paso del algoritmo las estimaciones son obtenidas según el paso M de la siguiente manera:

$$\hat{\Gamma}^{(k+1)} = (\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}\tilde{\Theta}^{(k)} \tag{25}$$

y

$$\hat{\Sigma}^{(k+1)} = \frac{1}{\text{tr}(\mathbf{D})}\left(\text{E}((\Theta - \tilde{\Theta}^{(k)})^T\mathbf{D}(\Theta - \tilde{\Theta}^{(k)})|\mathbf{X}, \mathbf{Y}, \Gamma^{(k)}, \Sigma^{(k)}) + (\Theta^{(k)} - \mathbf{X}\Theta^{(k)})^T\mathbf{D}(\Theta^{(k)} - \mathbf{X}\Theta^{(k)})\right) \tag{26}$$

El paso E de la k-ésima iteración consiste en encontrar solución a las siguientes esperanzas *a posteriori* necesarias para avanzar en el cálculo de las ecuaciones (25) y (26):

$$\tilde{\Theta}^{(k)} = \mathrm{E}(\Theta|\mathbf{X}, \mathbf{Y}, \Gamma^{(k)}, \Sigma^{(k)}) \tag{27}$$

y

$$\mathrm{E}(\Theta^T\mathbf{D}\Theta|\mathbf{X}, \mathbf{Y}, \Gamma^{(k)}, \Sigma^{(k)}) \tag{28}$$

La evaluación de esas dos últimas cantidades puede ser realizada por dos algoritmos: integración por cuadratura o aproximación de Laplace.

Para hacer una revisión exhaustiva del método, incluso de la convergencia y de los valores iniciales el lector puede referirse a Rubin (1991).

2.4. Valores plausibles

Para encontrar las estimaciones de los parámetros de la regresión latente, el algorito EM puede ser resumido como sigue:

1. Especificar o encontrar los valores iniciales $\Gamma^{(0)}$ y $\Sigma^{(0)}$ por medio de estimación vía máxima verosimilitud usando cuadraturas.
2. Paso E: evaluar la media y la varianza de la distribución *a posteriori* del vector de rasgo latente Θ como en las expresiones (27) y (28). Esto se puede abordar por cuadraturas o por aproximación de Laplace, dependiendo de la dimensión del rasgo latente. En modelos unidimensionales la integración puede hacerse por cuadraturas.
3. Paso M: actualizar los valores de Γ y Σ como en las expresiones (25) y (26)
4. Por medio de un criterio adecuado evaluar la convergencia. En el caso de que sea rechazada, volver al paso 2. En el otro caso definir $\hat{\Gamma}$ y $\hat{\Sigma}$.

En cuanto $\hat{\Gamma}$ y $\hat{\Sigma}$ sean encontrados, es posible definir un algoritmo de estimación aleatoria para R valores plausibles y para cada unidad de la población como sigue:

1. Seleccionar de manera aleatoria $\Gamma^{(r)}$ de la distribución $N(\hat{\Gamma}, \hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{D}\mathbf{X})$
2. Calcular la media de la distribución asumida para θ_i como $x_i^T\Gamma^{(r)}$ con x_i^T el vector fila de la matriz $\mathbf{X}$.
3. Usando $\hat{\Sigma}$ como la matriz de varianzas de la distribución *a priori* $P(\theta|\mathbf{X})$ de θ, determinar la distribución *a posteriori* del rasgo latente θ mostrada en la ecuación (20) haciendo uso además del modelo de respuesta al ítem.
4. Seleccione de manera aleatoria un valor imputado $\tilde{\theta}_i^{(r)}$ de la distribución *a posteriori* encontrada en el número 3.

5. Repita el proceso para $r = 1, 2, \ldots, R$.

Otros mecanismos de selección aleatoria pueden ser usados. En particular, además de este, el ETS menciona otro en su *software* DESI (*Direct Estimation Software Interactive*) como sigue:

1. Calcular la media de la distribución asumida para θ_i como $x_i^T \hat{\Gamma}$ con x_i^T el vector fila de la matriz $\mathbf{X}$.
2. Usando $\hat{\Sigma}$ como la matriz de varianzas de la distribución a priori $P(\theta|\mathbf{X})$ de θ, determinar la distribución *a posteriori* del rasgo latente θ mostrada en la ecuación (20) haciendo uso además del modelo de Respuesta al Ítem.
3. Seleccione de manera aleatoria un valor imputado $\tilde{\theta}_i^{(r)}$ de la distribución a posteriori encontrada en el número 2.
4. Repita el proceso para $r = 1, 2, \ldots, R$.

En general se usa $R = 5$. El tratamiento estadístico de los valores imputados se basa en producir estimaciones de parámetros de interés en la siguiente vía:

1. Cada parámetro es estimado para cada uno de los R valores plausibles y todas las estimaciones se promedian.
2. El error estándar de este promedio estimado es calculado combinando el promedio del error de muestreo de las R estimaciones con la varianza de las R estimaciones

Es importante destacar que si las escalas sobre las que proyectan las mediciones de la población están bien construidas, la inclusión de cualquier variable de clasificación en la regresión, no debe sesgar el resultado de la imputación múltiple realizado sobre la distribución del rasgo latente.

3. Ejemplo: un caso de simulación

El objetivo de la simulación es observar el uso de valores plausibles como alternativa para corregir el sesgo y la imprecisión provocada por aplicar a cada individuo solo una muestra de los ítems. Para esto se consideran 4.000 estudiantes repartidos en 160 establecimientos educativos, y para cada uno de ellos se definen dos variables asociadas: sexo (masculino o femenino) relacionado a cada individuo y sector (oficial o no oficial) relacionado a cada establecimiento al que pertenece. La tabla (1) relaciona los promedios y las desviaciones inducidas para cada uno de los cuatro grupos poblacionales.

De los dos factores con que se simula la población, se induce que uno sea factor condicionante (el sector) y el otro no (sexo). Se aprecia que la diferencia de resultados entre hombres y mujeres es nula, mientras que la diferencia entre agregaciones

Tabla 1: *Promedio y desviación estándar usado para generar la simulación. Fuente: elaboración propia.*

	Masculino	Femenino	Promedio
Oficial	-0.707(0.707)	-0.707(0.707)	-0.707(0.707)
No oficial	0.707(0.707)	0.707(0.707)	0.707(0.707)
Promedio	0.000(0.707)	0.000(0.707)	0.000(0.707)

de sector oficial y no oficial es de 1.414 unidades, es decir, aproximadamente $\sqrt{2}$, cada una de ellas con una desviación estándar aproximadamente de 1.

Por otro lado, se considera que hay 72 ítems y se seleccionan grupos de 12 de manera aleatoria para conformar 6 bloques: A, B, C, D, E y F. Un bloque es un grupo de preguntas que responde a las mismas características de contenido y de dificultad que la prueba completa.

Se consideran solamente seis cuadernillos (combinación de bloques) y se construyen tres escenarios de simulación así:

1. El escenario poblacional. Todos los estudiantes presentan toda la prueba, es decir, a cada estudiante se le suministran los 72 ítems en los 6 bloques.
2. Cada estudiante presenta cuatro bloques de seis, es decir, 48 ítems (4/6).
3. Cada estudiante presenta dos bloques de seis, es decir, 24 ítems (2/6).

El ejercicio consiste en simular un conjunto de datos para cada escenario bajo las mimas condiciones: el cambio del escenario 1 al 2 consiste en eliminar de las observaciones los dos últimos bloques y a su vez el cambio del escenario 2 al 3 consiste en eliminar de las observaciones los dos últimos bloques. Luego de simular bajo normalidad la habilidad de cada estudiante se simulan también bajo normalidad los valores de los parámetros que definen el modelo. Para el ejercicio se usa el modelo logístico de dos parámetros 2PL. Por último, como insumo de calibración y calificación, es necesario simular el vector de respuestas $\mathbf{Y}$.

Sea $\tilde{\theta}_i$ la habilidad simulada para el individuo i y sean $\tilde{a}_j$ y $\tilde{b}_j$ los parámetros de discriminación y dificultad simulados para el ítem j. Los siguientes pasos son usados para generar la variable Y_{ij} que sigue una distribución Bernoulli condicional con $p_{ij} = P(Y_{ij} = 1|\theta_i)$:

1. Se genera la cantidad u con $U \sim U(0, 1)$
2. Se calcula
$$\tilde{p}_{ij} = P(Y_{ij} = 1|\tilde{\theta}_i) = \frac{1}{1 + \exp\{-\tilde{a}_j(\tilde{\theta}_i - \tilde{b}_j)\}} \tag{29}$$
3. Se evalúa $u < \tilde{p}_{ij}$. Si es verdadero se asigna $Y_{ij} = 1$. En caso contrario se asigna $Y_{ij} = 0$.

Escenario 1: poblacional

En este escenario cada estudiante presenta todos los bloques. En el caso de que cada estudiante presente todos los ítems de la prueba, no es necesario imputar su habilidad promedio de acuerdo con algún agregado, es decir, no es necesario usar valores plausibles y el propósito de la prueba será distinto. El diseño puede mostrarse como en la tabla (2):

Tabla 2: *Diseño de cuadernillos en el escenario 1. Fuente: elaboración propia.*

Cuadernillo	Bloq. 1	Bloq. 2	Bloq. 3	Bloq. 4	Bloq. 5	Bloq. 6
1	A	B	C	D	E	F
2	F	A	B	C	D	E
3	E	F	A	B	C	D
4	D	E	F	A	B	C
5	C	D	E	F	A	B
6	B	C	D	E	F	A

Este escenario es llamado poblacional, ya que asume que toda la población de estudiantes se mide en cada uno de los ítems que componen la prueba y sirve como referencia para comparar los resultados de imputación múltiple cuando cada estudiante no presenta todos los ítems. Haciendo uso del vector $\mathbf{Y}$, se ajusta el modelo para cada parámetro y se compara contra los valores inducidos en la simulación. La figura (1) muestra en la parte izquierda para cada uno de los 72 ítems la dificultad simulada contra la dificultad estimada y en la parte derecha la distribución de los 72 ítems de acuerdo con su dificultad simulada o estimada.

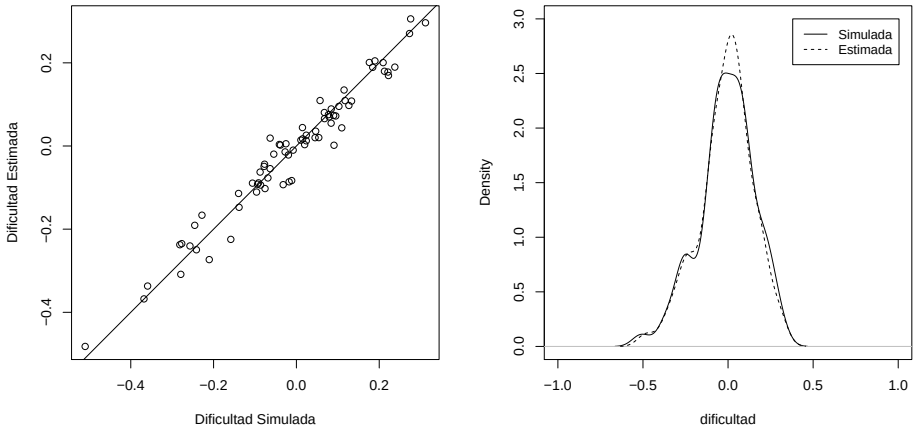

Figura 1: *a). Diagrama de dispersión de la dificultad simulada y estimada / b). Distribución de la dificultad simulada y estimada en los 72 ítems. Fuente: elaboración propia.*

La figura muestra que el ajuste del parámetro de dificultad del modelo es bueno y que, salvo por la variabilidad habitual, la asociación entre lo simulado y lo observado es evidente. La figura (2) muestra a izquierda, para cada uno de los 4000 individuos, la habilidad simulada versus la habilidad estimada y, a derecha muestra la distribución de los 4000 individuos de acuerdo con su habilidad simulada o estimada.

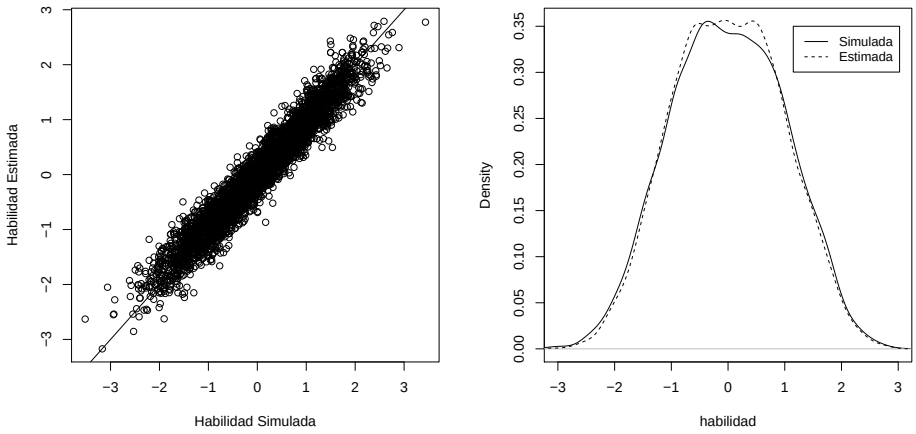

Figura 2: *a). Diagrama de dispersión de la habilidad simulada y estimada / b). Distribución de la habilidad simulada y estimada en los 4000 individuos. Fuente: elaboración propia.*

Se observa que el ajuste de la habilidad de cada individuo por parte del modelo es bueno y la asociación entre lo simulado y lo observado es importante. La figura 3 relaciona la densidad de la distribución de las habilidades de la población en general y cuando se agrega por la variable del sector del establecimiento educativo. El estimador usado es el puntaje esperado *a posteriori* (PEP).

Se observa que el sector oficial muestra un rasgo latente promedio más bajo que el sector no oficial. Esas cantidades son -0.6651 y 0.6674 respectivamente, con una desviación estándar de 0.708 y 0.710. Se encuentra además que la variable sexo no es significativa, ya que en promedio hombres y mujeres obtienen una habilidad promedio de -0.003 y 0.004. Las magnitudes son muy cercanas a lo inducido por la simulación, pero no exactas, por lo que este resultado es el punto de referencia de lo que se quiere medir cuando se deja de observar información, como en el caso de los dos siguientes escenarios.

Escenario 2: diseño de 4 bloques de 6

En este escenario cada estudiante presenta 4 bloques de 6. La tabla (3) relaciona el diseño.

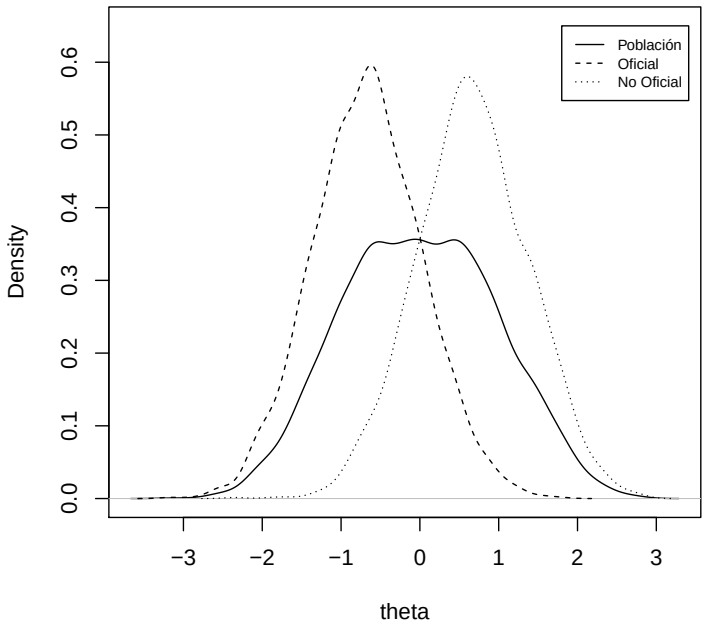

Figura 3: *Distribución de la habilidad en la población y por sector. Fuente: elaboración propia.*

Tabla 3: *Diseño de cuadernillos en el escenario 2. Fuente: elaboración propia.*

Cuadernillo	Bloq. 1	Bloq. 2	Bloq. 3	Bloq. 4
1	A	B	C	D
2	F	A	B	C
3	E	F	A	B
4	D	E	F	A
5	C	D	E	F
6	B	C	D	E

De la tabla 3 se observa que 4 de 6 estudiantes presentan cada bloque, es decir, el 2/3 de la población. Bajo este escenario se presenta el problema de no observar toda la información de la prueba en cada estudiante. La figura 4 muestra en la parte izquierda para cada uno de los 72 ítems la dificultad simulada contra la dificultad estimada con la característica que no usa toda la población; en la parte derecha muestra la distribución de los 72 ítems de acuerdo con su dificultad simulada y estimada bajo dichas condiciones.

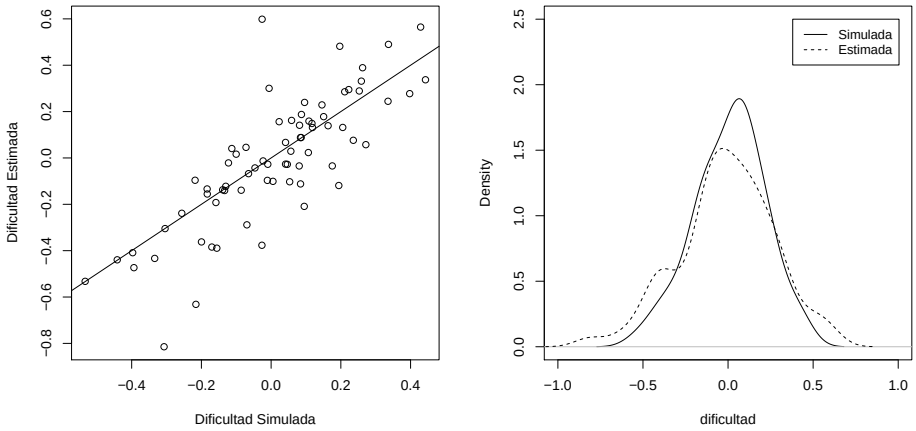

Figura 4: *a). Diagrama de dispersión de la dificultad simulada y estimada (4/6) / b). Distribución de la dificultad simulada y estimada (4/6) en los 72 ítem. Fuente: elaboración propia.*

La figura muestra que el ajuste del parámetro de dificultad del modelo es bueno aunque la asociación entre lo simulado y lo observado se distorsiona en comparación con lo observado en la figura del escenario poblacional (1). Como se espera en el caso en que se observa menos información de la esperada, las estimaciones del modelo presentan pérdida de precisión y aumenta el riesgo de incluir sesgos. La figura 5 muestra en la parte izquierda la habilidad simulada de cada uno de los 4000 individuos contra la habilidad estimada bajo este escenario y en la parte derecha muestra la distribución de los 4000 individuos de acuerdo con su habilidad simulada y estimada.

La figura muestra que el ajuste de la habilidad de cada individuo por parte del modelo es bueno aunque la asociación entre lo simulado y lo observado adquiere una variabilidad destacada. Es importante señalar que al perder información en cada individuo, la varianza de la habilidad es subestimada. Una de las implicaciones directas que tiene este tipo de fenómeno es que se puede inflar el error de tipo I, ocasionando eventualmente declarar diferencias en algunos grupos subpoblacionales cuando dichas diferencias realmente no existen. La figura (6) muestra la densidad de la distribución de las habilidades de la población en general y por sector del establecimiento educativo. Al igual que en el caso anterior, la habilidad del estudiante es estimada por medio del estimador bayesiano.

El sector oficial muestra un rasgo latente promedio más bajo que el sector no oficial, manteniendo la relación de magnitudes con que se indujo la simulación. Sin embargo estas cantidades son -0.4573 y 0.4532 respectivamente, con una desviación estándar de 0.665 y 0.664. Al igual que el caso anterior, se encuentra que la

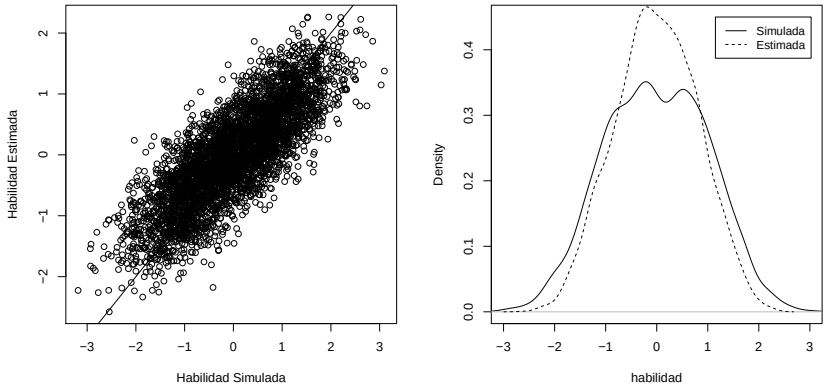

Figura 5: *a). Diagrama de dispersión de la habilidad simulada y estimada (4/6) / b). Distribución de la habilidad simulada y estimada (4/6) en los 4000 individuos. Fuente: elaboración propia.*

variable sexo no es significativa, ya que en promedio hombres y mujeres obtienen una habilidad promedio de -0.008 y 0.008. Las relaciones son muy cercanas a lo inducido por la simulación y a lo observado en el caso poblacional, pero las magnitudes muestran cambios importantes. Esto es debido a que la prueba no es suministrada en su totalidad a los estudiantes y, cada uno de ellos presenta solamente una parte de la misma. Así se evidencia que el agregado del puntaje esperado *a posteriori* es un estimador sesgado del promedio.

Escenario 3: diseño de 2 bloques de 6

En este escenario cada estudiante presenta 2 bloques de 6. El diseño se presenta en la tabla (4).

Tabla 4: *Diseño de cuadernillos en el escenario 3. Fuente: elaboración propia.*

Cuadernillo	Bloq. 1	Bloq. 2
1	A	B
2	F	A
3	E	F
4	D	E
5	C	D
6	B	C

De la tabla (4) se observa que 2 de 6 estudiantes presentan cada bloque, es decir, 1/3 de la población. La figura (7) relaciona en la parte izquierda la dificultad simulada de cada uno de los 72 ítems contra la dificultad estimada y en la parte derecha muestra la distribución de los 72 ítems de acuerdo con su dificultad simulada y

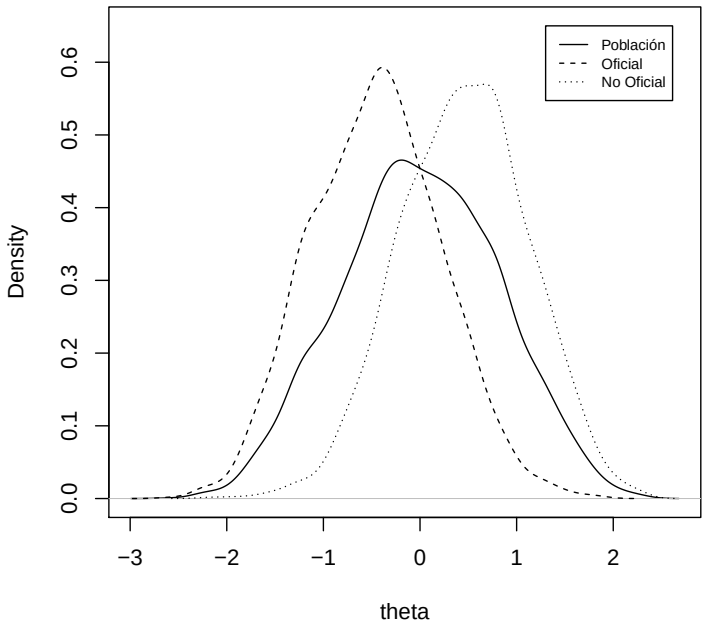

Figura 6: *Distribución de la habilidad en la población y por sector. Fuente: elaboración propia.*

estimada bajo dichas condiciones.

La figura muestra que el ajuste del parámetro de dificultad del modelo es bueno aunque la asociación entre lo simulado y lo observado se distorsiona más que el escenario inmediatamente anterior. La figura 8 muestra en la parte izquierda la habilidad simulada para cada uno de los 4000 individuos contra la habilidad estimada y a derecha muestra la distribución de los 4000 individuos de acuerdo a su habilidad simulada y estimada.

Al perder información en cada individuo, la varianza de la habilidad se subestima. La figura (9) muestra la densidad de la distribución de las habilidades de la población y cuando se agrega por la variable del sector del establecimiento educativo.

El sector oficial muestra un rasgo latente promedio más bajo que el sector no oficial, manteniendo la relación de magnitudes con que se indujo la simulación y que se ha observado en los dos escenarios mencionados anteriormente. Las magnitudes en este escenario son -0.3690 y 0.3476 respectivamente, con una desviación estándar de 0.633 y 0.631. En promedio hombres y mujeres obtienen una habilidad de -0.005 y 0.005. Con respecto al escenario anterior, es fácil deducir que el resultado

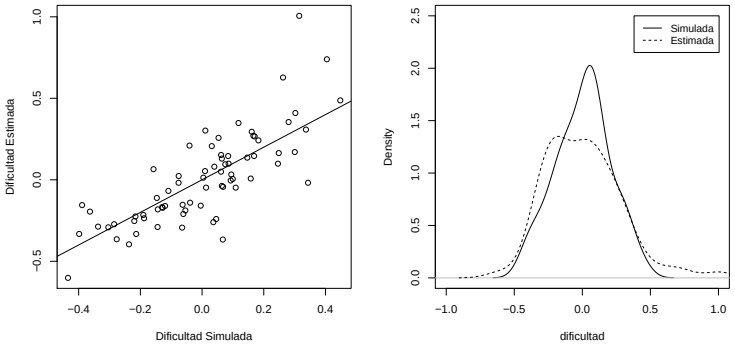

Figura 7: *a). Diagrama de dispersión de la dificultad simulada y estimada (2/6) / b). Distribución de la dificultad simulada y estimada (2/6) en los 72 ítem. Fuente: elaboración propia.*

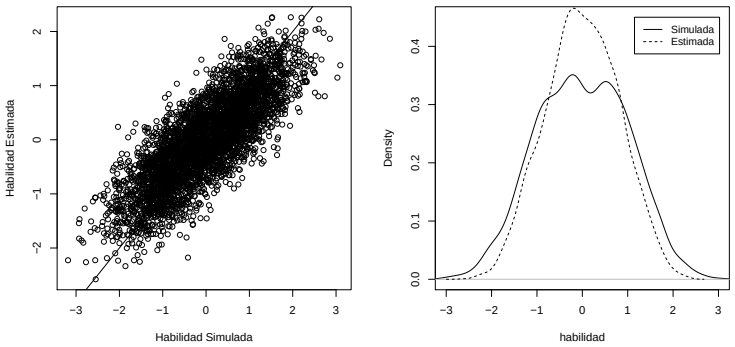

Figura 8: *a). Diagrama de dispersión de la habilidad simulada y estimada (2/6) / b). Distribución de la habilidad simulada y estimada (2/6) en los 4000 individuos. Fuente: elaboración propia.*

agregado en este tiende a minimizar las diferencias entre los promedios agregados de los establecimientos oficiales y no oficiales.

Escenario 4: valores plausibles para un diseño de 2 bloques de 6

Se observa que en cuanto el número de ítems evaluados por individuo decrece, los resultados agregados según algunas variables de interés sufren un cambio trascendental que no corresponde con la realidad. La figura (10) muestra el cambio de la distribución de las habilidades estimadas de todos los individuos que pertenecen a instituciones oficiales.

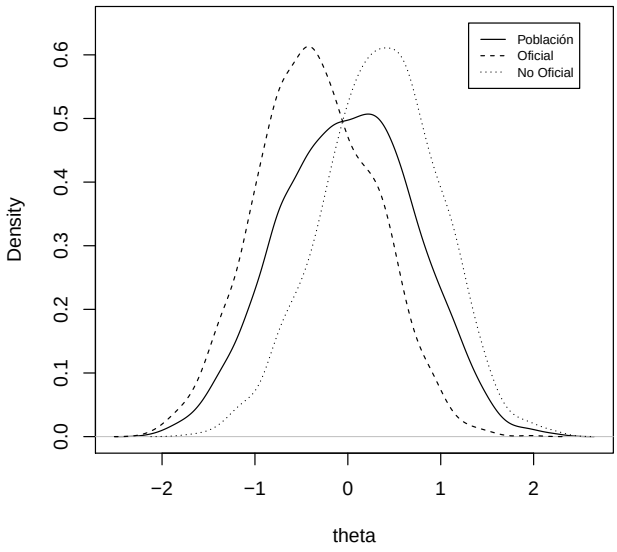

Figura 9: *Distribución de la habilidad en la población y por sector. Fuente: elaboración propia.*

De la figura se observa que aparece un sesgo importante hacia la derecha. Esto es debido además al estimador usado, pues el estimador bayesiano es levemente sesgado hacia el promedio poblacional que en este caso es 0. Para corregir esto, es necesario hacer el procedimiento de imputación múltiple. En este ejercicio se escogen cinco valores plausibles de la distribución *a posteriori* mencionada en (17) para cada individuo y se calculan las estadísticas de interés como en la sección (2.3). La figura (11) asocia el cambio de la distribución de la habilidad en este grupo poblacional en los tres escenarios mencionados junto con el de los valores plausibles.

Se observa que cuando el escenario es el de los valores plausibles, la distribución se acerca más a la poblacional, corrigiendo el sesgo y la variabilidad sobre todo que pierde en el escenario de los dos bloques aplicados de seis. Numéricamente los resultados promedio son mostrados en las tablas (5 y 6).

Con la tabla se puede verificar que existe pérdida de variabilidad observada cuando no se aplican todos los ítems a la población completa. Sin embargo, el método de los valores plausibles ayuda a corregir esa variabilidad y ayuda en este caso específico, a la corrección del sesgo.

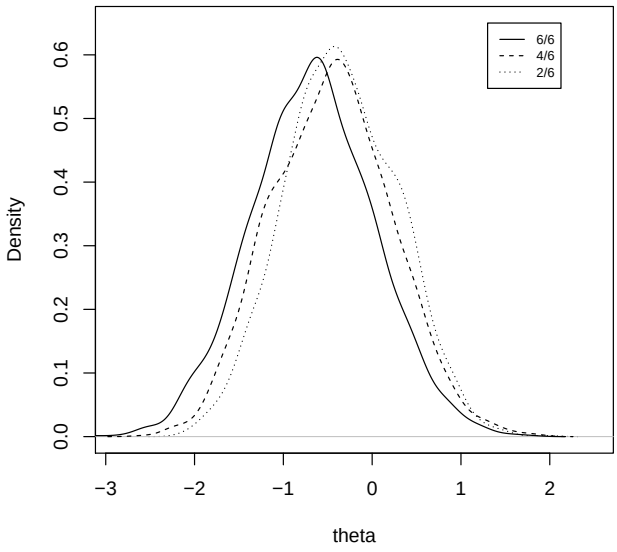

Figura 10: *Distribución de la habilidad de los estudiantes del sector oficial en los tres escenarios. Fuente: elaboración propia.*

Tabla 5: *Media y desviación por sector y zona en el escenario teoríco y poblacional. Fuente: elaboración propia.*

Sexo	Sector	Teórico	Poblacional (6/6)
M	O	-0.71 (0.71)	-0.6651 (0.6922)
M	N	0.71 (0.71)	0.6585 (0.7324)
F	O	-0.71 (0.71)	-0.6672 (0.69)
F	N	0.71 (0.71)	0.6764 (0.6787)

Tabla 6: *Media y desviación por sector y zona en los escenarios (4/6) y(2/6). Fuente: elaboración propia.*

Sexo	Sector	Escenario (4/6)	Escenario (2/6)	Escenario (2/6 - VP)
M	O	-0.4573 (0.6653)	-0.369 (0.6338)	-0.6842 (0.7411)
M	N	0.4396 (0.658)	0.3575 (0.6105)	-0.6226 (0.7541)
F	O	-0.4494 (0.6857)	-0.3266 (0.6279)	0.6861 (0.7466)
F	N	0.4667 (0.6706)	0.3377 (0.6508)	0.6252 (0.7514)

4. Conclusiones

Las pruebas estandarizadas de gran escala usan bloques aleatorizados para garantizar la cobertura necesaria que impone el marco de referencia definido y para

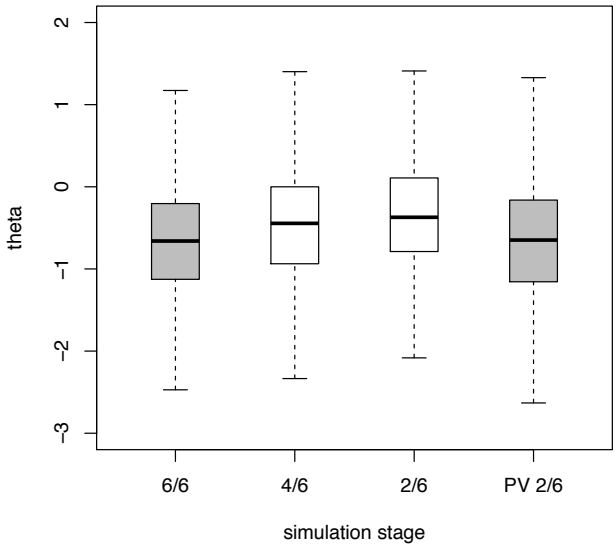

Figura 11: *Distribución de la habilidad de los estudiantes del sector oficial en los cuatro escenarios. Fuente: elaboración propia.*

garantizar la economía y practicidad del operativo de la prueba en campo. Esto implica una pérdida de información considerable que debe ser tratada con cuidado. El no tratamiento adecuado de este problema podría traer consecuencias graves en las estadísticas agregadas y en la toma de decisiones con respecto a la población estudiada.

El estimador bayesiano usado en este ejercicio para calcular el rasgo latente de cada individuo es ligeramente sesgado hacia la media poblacional de su distribución. En el ejercicio se aprecia que dicho sesgo crece en cuanto el total de bloques que se mide en cada individuo disminuye. Además de perder información que se debe observar, la variabilidad sobre los resultados de la distribución del rasgo latente en la población decrece. Esto puede provocar inferencias erróneas, aumentando la probabilidad de cometer error tipo I.

El método de los valores plausibles es un caso específico de imputación de la información por medio de la inclusión de variables de clasificación de los individuos para la generación de una distribución *a posteriori* del rasgo latente. El método ayuda a controlar la pérdida de la variabilidad mencionada, y en el caso particular de este estimador, corregir el sesgo obtenido.

Recibido: 22 de febrero del 2016
Aceptado: 4 de abril del 2016

Referencias

Aitkin, M. & Aitkin, I. (2011), *Statistical Modeling of the National Assessment of Educational Progress*, Nueva York: Springer.

Bock, R. D. & Aitkin, M. (1981), 'Marginal maximum likelihood estimation of item parameters: Application of an em algorithm', *Psychometrika* **46**(4), 443–459.

Bock, R. D. & Mislevy, R. J. (1982), 'Adaptive EAP estimation of ability in a microcomputer environment', *Applied psychological measurement* **6**(4), 431–444.

González, E. & Rutkowski, L. (2010), 'Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments', *IERI monograph series: Issues and methodologies in large-scale assessments* .

Harwell, M. R., Baker, F. B. & Zwarts, M. (1988), 'Item parameter estimation via marginal maximum likelihood and an em algorithm: A didactic.', *Journal of Educational and Behavioral Statistics* .

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983), *Item response theory: Application to psychological measurement.*, Dow Jones-Irwin, Homewood, IL.

Informe técnico SABER 5o. y 9o. 2009 (n.d.), Technical report.

Kass, R. & Steffey, D. (1989), 'Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models)', *Journal of the American Statistical Association* **84**(407), 717–726.

Rubin, D. B. (1991), 'EM and beyond', *Psychometrika* **56**(2), 241–254.