UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Trabajo de grado
Septiembre 2018

Propuesta metodológica para el aumento de datos en muestras pequeñas para la estimación de parámetros de ítems

Methodological Proposal for Data Augmentation in Small Samples for Item Parameters Estimation

Jeison Enrique Rodríguez García[a]
jeisonrodriguez@usantotomas.edu.co

Michel Felipe Córdoba Perozo[b]
mfcordobap@unal.edu.co

Resumen

En la teoría de respuesta al ítem (TRI) se han estudiado numerosos factores que influyen en la precisión de la estimación de los parámetros de los ítems, entre ellos, la cantidad de ítems que componen el test, la cantidad de individuos que se evalúan, el método de estimación empleado, la distribución de las habilidades ajustadas de los individuos, entre muchos otros. Se conoce que la cantidad de individuos evaluados es uno de los factores que más influye en la precisión de la estimación de dichos parámetros dependiendo del modelo empleado (Sahin & Anil 2016). Las metodologías de estimación de parámetros usadas en la TRI requieren una cantidad mínima de individuos evaluados para obtener estimaciones precisas de los parámetros de los ítems (Hambleton 1989).

Sin embargo, estas cantidades no siempre pueden ser alcanzadas por diversas razones, como por ejemplo que la población evaluada no es suficientemente grande para alcanzar la muestra requerida o que por temas de presupuesto resulta muy costoso acceder a una muestra mínima. El objetivo de este trabajo consiste en proponer e implementar una imputación múltiple con valores plausibles en una técnica llamada DuPER por sus siglas en inglés (Duplicate, Erase, Replace) que es una técnica de aumento de datos (data augmentation) cuya finalidad es expandir una muestra de individuos para obtener información suficiente para realizar estimaciones precisas de parámetros de ítems.

–

***Palabras clave*:** Teoría de respuesta al ítem (TRI), aumento de datos, DuPER, valores plausibles, estimación de parámetros de ítems, precisión en la estimación.

Abstract

In the item response theory (IRT), numerous factors have been studied that influence the accuracy of the item parameter estimation, including the number of items that make up the test, the number of individuals that are evaluated, the estimation method used, the distribution of the adjusted scores of the individuals, among many others. It is known that the number of individuals evaluated is one of the factors that most influence the accuracy of the estimation of these parameters depending on the model used (Sahin & Anil 2016). The parameter estimation methodologies used in the IRT require a minimum number of individuals evaluated to obtain accurate estimates of the items parameters (Hambleton 1989).

However, these amounts can not always be reached for various reasons, such as for example that the population evaluated is not large enough to reach the required sample or that due to budget issues it

[a]Estudiante
[b]Director

is very expensive to access a minimum sample. The objective of this work consists of to propose and implement a multiple imputation with plausible values in a technique called DuPER for its acronym (Duplicate, Erase, Replace) which is a technique of data augmentation whose purpose is to expand a sample of individuals to obtain sufficient information to make precise estimates of items parameters.

_

***Keywords*:** Item Response Theory (IRT), data augmentation, DuPER, plausible values, item parameter estimation, estimation accuraccy.

1. Introducción

En la actualidad, las pruebas estandarizadas o test tienen un extenso uso en el sector educativo, médico, industrial, entre muchos otros, con el fin de realizar un diagnóstico (test no cognitivos diseñados para medir intereses, actitudes y otros aspectos no cognitivos) o una clasificación cognitiva (test cognitivos que evalúan las habilidades de los individuos) (Fox 2010). Para evaluar o analizar las respuestas de los test bajo condiciones estandarizadas, se usan modelos de teoría de respuesta al ítem que asignan un puntaje a cada individuo. Los modelos de TRI en test cognitivos con ítems dicotómicos (dos posibles respuestas: correcta o incorrecta) son usados para calcular la probabilidad de conseguir una respuesta específica basándose en la habilidad del evaluado y las características del ítem.

Los modelos logísticos de dos o tres parámetros (2PL o 3PL respectivamente) requieren de una cantidad mínima de muestra (individuos evaluados) para tener estimaciones precisas de los parámetros de los ítems (de Ayala 2009). Teóricamente, la precisión de una estimación está ligada a su error estándar, si el tamaño de la muestra es pequeño, las estimaciones de los parámetros de los ítems tienen errores estándar muy grandes y al usarlos para estimar habilidades de los evaluados se induciría un sesgo (Linacre 1994).

En algunos casos y dependiendo del constructo o población que se desea evaluar con una prueba estandarizada o test, es difícil obtener los tamaños de muestra mínimos requeridos. En el contexto educativo colombiano este problema se presenta cuando se requiere evaluar poblaciones de estudiantes de programas específicos que no cuentan con una cantidad suficiente de individuos (programas técnicos), o cuando por las características del test (pilotajes de pruebas de inglés) resulta económicamente muy costosa su implementación.

En la literatura existe una metodología llamada aumento de datos (data augmentation) usada extensamente en el campo del machine learning, específicamente en el deep learning aplicado a imágenes cuando no se cuenta con suficientes datos. Esta metodología consiste en aplicar transformaciones (rotaciones, cambios de escala, recortes, etc) a la información existente (imágenes) con el fin expandirla sistemáticamente y así tener suficiente información.

Para las aplicaciones en pruebas estandarizadas y TRI no existen muchos antecedes del uso de técnicas de aumento de datos, Patrick Foley (2010) propone, desarrolla e implementa una técnica llamada DuPER en escenarios de simulación donde varía la cantidad de ítems, la cantidad de evaluados, cantidad de réplicas por evaluado, tasas de eliminación y los métodos de imputación: algoritmo EM y Cadenas de Marcov vía Monte Carlo (MCMC). En este estudio, Foley concluye que la aplicación de la técnica DuPER sobre la mayoría de los escenarios resulta en altos RMSE y bajas correlaciones para las estimaciones de los parámetros de los ítems. Además sugiere para trabajos futuros, implementar la técnica en datos reales, ya que los resultados de su estudio pueden estar sesgados por las distribuciones de los parámetros de los ítems y de las habilidades de los evaluados inducidas en las simulaciones.

En este trabajo se implementa la técnica DuPER en una muestra pequeña seleccionada aleatoriamente de una población de técnicos y tecnólogos evaluados en la prueba Saber TyT realizada por el Instituto Colombiano para la Evaluación de la Educación (ICFES), utilizando una imputación múltiple con valores

plausibles para ajustar el error estándar de las estimaciones de los parámetros. Se utilizan criterios de evaluación como el coeficiente de correlación de Pearson, RMSE (raíz del error cuadrático medio), Sesgo y coeficiente de regresión para probar la eficacia y precisión de dichas estimaciones.

En la sección 2: Modelos de teoría de respuesta al ítem (TRI), se describen los modelos generalmente usados en TRI para evaluar ítems dicotómicos y que se usarán en este trabajo, así como algunas metodologías para estimar parámetros de ítems. En la sección 3: Metodología, se describe la técnica de aumento de datos llamada DuPER, junto con la teoría de la técnica de imputación múltiple con valores plausibles y los criterios para evaluar la eficacia de los resultados obtenidos con la técnica propuesta, en la sección 4: Resultados se mostrarán los resultados obtenidos y en la sección 5: Conclusiones, se harán comentarios basados en los resultados y finalmente en la sección 6: Discusión se hablará sobre los hallazgos, limitaciones y recomendaciones para trabajos futuros.

2. Modelos de teoría de respuesta al ítem (TRI)

Los modelos de TRI definen una relación o correspondencia entre variables o rasgos latentes y sus manifestaciones. Estas metodologías usan caracterizaciones de los individuos o evaluados y de los ítems como predictores de los patrones de respuestas observadas en un test (de Ayala 2009). Estos modelos se enfocan en evaluar el comportamiento de los ítems individuales de un test en lugar de todo el conjunto simultáneamente. Para esto se usan funciones no lineales con el fin de relacionar las propiedades o parámetros de los ítems y las características de los evaluados con la probabilidad de dar una respuesta particular (Patrick Foley 2010); Matemáticamente esta relación se define como:

$$p(\theta_i) = p(\boldsymbol{X_{ij}} = x_{ij}|\theta_i, \delta_j). \tag{1}$$

Esta ecuación indica que la probabilidad de que el $i-$ésimo individuo responda $x_{ij} = \{1$ (respuesta correcta) o 0 (respuesta incorrecta)$\}$ al $j-$ésimo ítem del vector de respuestas $\boldsymbol{X}$ depende de la habilidad de dicho evaluado (θ_i) y de las características o parámetros del ítem (δ_j). El desarrollo de esta función matemática en general se basa en algunos supuestos (Reckase 2009), que para los modelos expuestos en este trabajo principalmente son:

- La habilidad de los individuos no cambia durante la implementación del test.
- Las características o parámetros de los ítems se mantienen constantes a las diferentes situaciones en las que se use el test (Reckase 2009).
- El supuesto de *unidimensionalidad*; establece que las observaciones de las respuestas de los ítems son una función de una única variable o rasgo latente del individuo (de Ayala 2009).
- El supuesto de *independencia condicional* o *independencia local* establece que la respuesta que un individuo le da a un ítem es independiente a las respuestas que le dio a cualquier otro ítem, esto condicionando a la habilidad del individuo (de Ayala 2009).

2.1. Modelos dicotómicos de TRI

La cantidad de categorías de respuesta de los ítems o la naturaleza de las mismas definen el tipo de modelo que se debe usar para encontrar la probabilidad de responder correctamente (para modelos dicotómicos) o la probabilidad de responder alguna categoría específica (para modelos politómicos). En este trabajo sólo se usan modelos dicotómicos y se detallan a continuación.

- **Rasch**

  Este modelo caracteriza cada ítem en términos de un sólo parámetro (δ_j) y representa su localización en la escala del rasgo latente que mide el constructo (de Ayala 2009). Este parámetro es usualmente llamado dificultad del $j-$ésimo ítem y se simboliza con b_j. En la escala mencionada se asume que los evaluados requieren una habilidad más alta para responder correctamente ítems con dificultad alta, que para responder correctamente ítems con dificultad baja.

  La distancia entre la ubicación del evaluado en la escala del rasgo latente y la ubicación de la dificultad del ítem ($\theta_i - b_j$) es un importante determinante de la probabilidad de respuesta correcta. Así el modelo que permite encontrar dicha probabilidad es:

$$p(x_{ij} = 1|\theta_i, b_j) = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}}. \tag{2}$$

  Donde $p(x_{ij} = 1|\theta_i, \delta_j)$ es la probabilidad de que el $i-$ésimo individuo responda correctamente, θ_i es su ubicación en el rasgo latente y b_j es la dificultad del ítem.

- **Logístico de un parámetro (1PL)**

  En el modelo logístico de un parámetro (1PL) además de la dificultad de los ítems, se tiene en cuenta una constante (α) que representa la capacidad discriminativa de todos los ítems. Esta constante permite ajustar mejor el modelo a la distribución de las probabilidades empíricas. Matemáticamente el modelo se expresa así:

$$p(x_{ij} = 1|\theta_i, \alpha, b_j) = \frac{e^{\alpha(\theta_i - b_j)}}{1 + e^{\alpha(\theta_i - b_j)}}. \tag{3}$$

  El modelo de Rasch es un caso particular del modelo 1PL cuando la constante de discriminación es igual a uno ($\alpha = 1$).

  Desde el punto de vista psicométrico, la diferencia entre el modelo de Rasch y el modelo 1PL es que el modelo 1PL está enfocado en ajustarse a los datos de la mejor forma posible, mientras que el modelo de Rasch es usado para reconstruir el constructo o rasgo latente de interés.

  Con este modelo, la suma de las respuestas correctas de un individuo (puntaje observado) es una estadística suficiente para la estimación de su localización en la escala del rasgo latente, y la suma de las respuestas correctas de un ítem j (puntaje del ítem) es una estadística suficiente para la estimación de la localización del ítem en la escala. Todos los evaluados que tengan la misma cantidad de respuestas correctas tendrán la misma estimación de su habilidad ($\hat{\theta}_i$), y todos los ítems que tengan la misma cantidad de respuestas correctas, tendrán la misma estimación de dificultad ($\hat{b}_j$). La precisión de $\hat{\theta}_i$ y $\hat{b}_j$ viene dada por sus respectivos errores estándar (de Ayala 2009).

- **Logístico de dos parámetros (2PL)**

  Una extensión del modelo 1PL es el modelo logístico de dos parámetros (2PL) que incorpora información sobre la capacidad discriminativa que tiene cada ítem (α_j) sobre la población de individuos o evaluados. A diferencia del modelo 1PL que utiliza una misma constante de discriminación para todos los ítems, en el modelo 2PL este parámetro varía para cada ítem j. En este modelo, la discriminación de cada ítem pondera la distancia entre la ubicación de los evaluados (habilidad) y la ubicación o dificultad del ítem ($\alpha_j(\theta_i - b_j)$). Matemáticamente el modelo incluyendo los dos parámetros de cada ítem se expresa así:

$$p(x_{ij} = 1|\theta_i, \alpha_j, b_j) = \frac{e^{\alpha_j(\theta_i - b_j)}}{1 + e^{\alpha_j(\theta_i - b_j)}}. \tag{4}$$

El parámetro de discriminación varía de -∞ a ∞ y valores entre 0.8 y 2.5 son considerados como razonablemente "buenos"(de Ayala 2009). Sí este parámetro toma valores negativos, indica que evaluados con habilidades bajas tienen una mayor probabilidad de responder correctamente un ítem que los evaluados con habilidades altas, por lo que dada la naturaleza y objetivo del test (diagnóstico cognitivo) se estaría cometiendo un error.

2.2. Métodos de estimación de parámetros de los ítems

Existen diferentes metodologías para estimar los parámetros de los ítems (δ_j) y la estimación el rasgo latente de cada individuo (θ_i). A continuación, se presentan brevemente dos de los métodos más usados en la estimación de los parámetros para ítems dicotómicos, empleando las respuestas de los ítems como mecanismo de estimación. Para ver más detalles de estas metodologías consulte (de Ayala 2009).

- **Máxima verosimilitud conjunta**

  Esta metodología maximiza la función de verosimilitud conjunta para estimar simultáneamente los parámetros de los ítems y las habilidades de los evaluados. Para encontrar la función de verosimilitud conjunta, se parte de la función de verosimilitud de los individuos, se debe asumir el supuesto de independencia condicional y que los ítems aplicados en el test son dicotómicos, así la probabilidad de que el $i-$ésimo individuo responda todo el test es el producto de las probabilidades de responder cada ítem, así:

$$p(\boldsymbol{X_i}|\theta_i, \delta) = \prod_{j=1}^{J} p_j^{x_{ij}} (1 - p_j)^{1-x_{ij}}. \tag{5}$$

  El término $p(\boldsymbol{X_i}|\theta_i, \delta)$ es la probabilidad de que el $i-$ésimo individuo responda el vector $\boldsymbol{X_i}$, condicionando a su habilidad (θ_i) y al vector de parámetros (δ) de los J ítems, $\delta = (\delta_1, ..., \delta_J)$. La probabilidad de responder correctamente un ítem j, es p_j y es calculada acorde a uno de los modelos expuestos anteriormente (por ejemplo 2PL).

  Para obtener la función de máxima verosimilitud conjunta L se parte de los parámetros de los J ítems y las habilidades de los N evaluados, para esto, se multiplica la expresión de la ecuación (5) para los N evaluados:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{J} p_j(\theta_i)^{x_{ij}} (1 - p_j(\theta_i))^{1-x_{ij}}, \tag{6}$$

  A partir de la ecuación (6) se puede inferir que a medida que aumenta la cantidad de ítems (J) o la cantidad de individuos evaluados (N) el producto de las probabilidades será potencialmente tan pequeño que será difícil de representar generando problemas de precisión numérica, para evitar esto, se suele realizar una transformación con logaritmo natural (log):

$$\log(L) = \sum_{i=1}^{N} \sum_{j=1}^{J} [x_{ij} \log(p_j(\theta_i)) + (1 - x_{ij}) \log(1 - p_j(\theta_i))]. \tag{7}$$

  Los valores de $\boldsymbol{\theta_i}$ y δ_j que maximizan la expresión anterior corresponden a las estimaciones finales. La estrategia de maximización de la función de verosimilitud conjunta consiste en una serie iterativa de pasos. En el primer paso los parámetros de los ítems se estiman usando unos valores provisionales (valores iniciales) de las habilidades de los individuos. Se hace de esta manera, porque usualmente se tienen muchos más individuos que ítems, por lo que se tendría más información para estimar los parámetros de los ítems. El segundo paso consiste en estimar las habilidades de los individuos

con las estimaciones de los parámetros de los ítems obtenidas en el paso anterior. Iterativamente se utilizan esas estimaciones de habilidades para volver a estimar los parámetros de los ítems de forma más precisa, y así generar nuevas estimaciones de habilidades. Este proceso se repite hasta que las estimaciones de los parámetros de los ítems y de las habilidades no cambie significativamente de una iteración a otra.

- **Máxima verosimilitud marginal**

Otra alternativa para realizar la estimación de los parámetros de los ítems y de las habilidades, es la máxima verosimilitud marginal (MML por sus siglas en inglés). En esta metodología, a diferencia de la máxima verosimilitud conjunta, se estiman los parámetros de los ítems y posteriormente usando máxima verosimilitud o alguna metodología bayesiana como la esperanza posterior, se estiman las habilidades de los individuos. Cuando se realiza la estimación de esta manera, teóricamente se tiene más precisión para algunos test. Así, por ejemplo, para test compuestos por 15 ítems o menos, la máxima verosimilitud conjunta produce un sesgo en la estimación de la habilidad de los individuos evaluados (de Ayala 2009).

La MML permite introducir información de la población de individuos para estimar los parámetros de los ítems sin tener directamente una estimación de sus habilidades. En este caso, dicha estimación es potencialmente dependiente de la distribución de la población.

Para introducir la función de MML, se tiene la probabilidad de que el $i-$ésimo individuo tenga el vector de respuestas $\boldsymbol{X_i}$:

$$p(\boldsymbol{X_i}|\theta_i, \delta) = \prod_{j=1}^{J} p_j^{x_{ij}} (1 - p_j)^{1-x_{ij}}. \tag{8}$$

donde la probabilidad de responder ese vector de respuestas está condicionada a la habilidad del individuo (θ_i) y a una matriz o vector de parámetros de los J ítems (δ). Matemáticamente la inclusión de la muestra de la población de individuos requiere una integración en la distribución de su población:

$$p(\boldsymbol{X_i}) = \int_{-\infty}^{\infty} p(\boldsymbol{X_i}|\theta_i, \delta) g(\theta_i|\upsilon) d\theta_i. \tag{9}$$

Aquí $g(\theta_i|\upsilon)$ representa la distribución continua de la habilidad del $i-$ésimo individuo, siendo υ un vector que contiene parámetros distribucionales de ubicación y escala de la población de individuos que generalmente son 0 y 1, respectivamente. Se debe notar que $p(\boldsymbol{X_i})$ ya no está condicionado a la habilidad y representa la probabilidad incondicional o marginal de que un evaluado seleccionado aleatoriamente de una población con distribución continua de habilidades $g(\theta_i|\upsilon)$ de un vector de respuestas $\boldsymbol{X_i}$.

Se asume que la distribución poblacional de las habilidades se conoce previamente o que tiene una forma previa, como una distribución normal por ejemplo. La integración anterior se puede aproximar usando una distribución discreta, dividida por $r = 1, ..., R$ puntos de corte en la distribución llamados puntos de cuadraturas (X_r) que ponderados por unos pesos $A(X_r)$ y sumados, se aproximan al área de la función usada (probabilidad). Aplicando los puntos de cuadraturas y sus pesos, el cálculo de la probabilidad de responder un vector $\boldsymbol{X_i}$, se simplifica a una suma ponderada de las probabilidades condicionales:

$$p(\boldsymbol{X_i})^* = \sum_{r=1}^{R} p(\boldsymbol{X_i}|X_r, \delta) A(X_r). \tag{10}$$

Para extender la expresión anterior a los N individuos en los J ítems, se parte de log verosimilitud de la función de verosimilitud marginal:

$$\log L = \sum_{i=1}^{N} \log p(\boldsymbol{X_i}) \tag{11}$$

Donde $p(\boldsymbol{X_i})$ está dado por la ecuación (10). Sustituyendo y simplificando en la ecuación (11) se obtiene la ecuación de verosimilitud marginal que se debe resolver para estimar un parámetro, en este caso la dificultad (b_j) del j-ésimo ítem en función de los patrones de respuesta individuales:

$$\frac{\partial}{\partial b_j} \log(L) = -\sum_{i=1}^{N} \int [x_{ij} - p_j(\theta_i)][p_j(\theta_i|x_{ij}, \delta, \upsilon)] d\theta_i, \tag{12}$$

donde $p_j(\theta_i|x_{ij}, \delta, \upsilon)$ está dado por

$$p_j(\theta_i|x_{ij}, \delta, \upsilon) = \frac{p(x_{ij}|\theta_i, \delta) g(\theta_i|\upsilon)}{p(x_{ij})}. \tag{13}$$

Expresando la ecuación (12) en términos de cuadraturas se tiene:

$$\frac{\partial}{\partial b_j} \log(L) = -\sum_{r=1}^{R} \sum_{i=1}^{N} [x_{ij} - p_j(X_r)][p_j(X_r|x_{ij}, \delta, \upsilon)], \tag{14}$$

y la ecuación (13) en términos de cuadraturas se tiene:

$$p_j(X_{r'}|x_i, \delta, \upsilon) = \frac{L(X_{r'})A(X_{r'})}{\sum_{r=1}^{R} L(X_r)A(X_r)}, \tag{15}$$

Donde $L(X_{r'}) = \prod_{j=1}^{J} p_j(X_{r'})^{x_{ij}} (1 - p_j(X_{r'}))^{1-x_{ij}}$ es una aproximación de la función de verosimilitud usando cuadraturas, y r' representa específicamente a alguno de los R puntos de cuadratura. Realizando simplificación de términos en la ecuación (14) (ver de Ayala 2009), se tiene que la ecuación de verosimilitud marginal con la cual se estiman los parámetros de los ítems es:

$$-\sum_{r'=1}^{R} [\bar{c}_{r'j} - \bar{n}_{r'j} p_j(X_{r'})] = 0. \tag{16}$$

Con $\bar{c}_{r'j} = \sum_{i=1}^{N} x_{ij} p(X_{r'}|x_{ij}, \delta, \upsilon) = \sum_{i=1}^{N} \frac{x_{ij} L(X_{r'})A(X_{r'})}{\sum_{r=1}^{R} L(X_r)A(X_r)}$ que refleja el número esperado de respuestas correctas para el j-ésimo ítem en cada uno de los puntos de cuadratura $X_{r'}$.

$\bar{n}_{r'j} = \sum_{i=1}^{N} p_j(X_r|x_{ij}, \delta, \upsilon) = \sum_{i=1}^{N} \frac{L(X_{r'})A(X_{r'})}{\sum_{r=1}^{R} L(X_r)A(X_r)}$ que es el número esperado de evaluados en cada punto de cuadratura $X_{r'}$.

La ecuación (16) tiene la forma de un puntaje de ítem observado menos un puntaje de ítem estimado. Las estimaciones de los parámetros corresponden a los valores que minimizan la suma de estas diferencias entre los R puntos de cuadratura.

La metodología de estimación de los parámetros de los ítems utilizada en este trabajo consiste en la implementación de un proceso que combina el método de estimación de parámetros de ítems MML y el algoritmo EM.

Los parámetros de los ítems se pueden considerar como datos faltantes provenientes de una distribución f parametrizada por ϕ. El objetivo del algoritmo EM es encontrar ϕ tal que la verosimilitud de $f(\cdot|\phi)$ es maximizada (Dempster & Rubin 1977). Así, la esencia del algoritmo EM es que a partir de un proceso iterativo de esperanza y maximización se refinan las estimaciones de los parámetros de los ítems, hasta un grado de precisión definido por un criterio de convergencia.

Los pasos efectuados en el proceso de estimación de parámetros son:

1. Calcular $\bar{n}_{r'j}$ y $\bar{c}_{r'j}$ de la ecuación (16) en la etapa de esperanza (etapa E). Si es la primera iteración, se toman valores iniciales aleatorios.
2. Los valores obtenidos en el paso 1 son usados para estimar los parámetros de los ítems con la función de verosimilitud marginal en la etapa de maximización (etapa M).
3. Las estimaciones obtenidas en el paso 2 son comparadas con los obtenidos en el paso 1.
    - Si la diferencia entre los dos conjuntos de estimaciones es mayor a un criterio de convergencia, se debe volver a ejecutar el proceso.
    - Si la diferencia entre los dos conjuntos de estimaciones es menor a un criterio de convergencia, las estimaciones finales son las obtenidas en el paso 2.

Para realizar estimaciones precisas de los parámetros de los ítems en un test, específicamente con un modelo 2PL, es necesario tener una cantidad mínima de evaluados (Hambleton 1989). Sin embargo, en ciertas ocasiones es difícil alcanzar la cantidad de evaluados deseada, ya sea por condiciones logísticas (difícil acceso o costos muy elevados) o porque simplemente no hay población suficiente. Si no se tiene una cantidad mínima de individuos evaluados en una muestra, es teóricamente inapropiado la aplicación de las metodologías de TRI para estimar parámetros de ítems.

3. Metodología

En módulos específicos de la prueba Saber TyT o de la prueba Saber PRO realizadas por el ICFES, podría presentarse el problema de no tener suficientes individuos en la población para realizar estimaciones de los parámetros de los ítems con modelos de TRI. En esta sección se introduce un método de aumento de datos llamado DuPER por sus siglas en inglés (Duplicate, Erase, Replace) (Patrick Foley 2010), con el fin de obtener información suficiente a partir de una muestra pequeña para tener estimaciones de parámetros de ítems precisas.

3.1. Técnica DuPER

En el contexto de las pruebas estandarizadas, "aumento de datos"se refiere al proceso mediante el cual se agregan vectores de respuesta plausibles a una muestra de evaluados. La técnica DuPER consiste en tres etapas: Replicar (Duplicate), Eliminar (Erase) y Reemplazar (Replace). Específicamente la técnica funciona de la siguiente manera (Patrick Foley 2010):

Dado un conjunto de evaluados (A, B y C) con sus respectivos vectores o strings de respuestas para 10 ítems

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	1	1	0	1	0	0	0	1	1	1
B	1	0	1	0	1	0	1	1	0	0
C	1	1	1	1	0	0	1	1	1	0

Tabla 1: Ítems (categorías 0 y 1)

1. El primer paso consiste en replicar cada vector de respuestas (o cada individuo) cierta cantidad de veces. Por ejemplo, si un test cuenta con tres participantes (A, B y C), cada vector de respuestas es replicado por decir tres veces, así que ahora se cuenta con un nuevo conjunto de datos con 9 evaluados y sus respectivos vectores de respuestas, llamados ahora pseudo evaluados y pseudo vectores de respuesta.

	1	2	3	4	5	6	7	8	9	10
A1	1	1	0	1	0	0	0	1	1	1
A2	1	1	0	1	0	0	0	1	1	1
A3	1	1	0	1	0	0	0	1	1	1
B1	1	0	1	0	1	0	1	1	0	0
B2	1	0	1	0	1	0	1	1	0	0
B3	1	0	1	0	1	0	1	1	0	0
C1	1	1	1	1	0	0	1	1	1	0
C2	1	1	1	1	0	0	1	1	1	0
C3	1	1	1	1	0	0	1	1	1	0

Tabla 2: Vectores de respuesta replicados tres veces

2. El siguiente paso consiste en eliminar observaciones (respuestas de los ítems) en cada uno de los pseudo vectores de respuestas con una razón o proporción definida previamente. Continuando el ejemplo, el test está compuesto por 10 ítems y usando procesos aleatorios, las respuestas de estos ítems para cada pseudo evaluado son eliminadas con una tasa de por ejemplo del 40 %. Con estas condiciones cada uno de los pseudo evaluados tendría 4 valores faltantes en su vector de respuestas.

	1	2	3	4	5	6	7	8	9	10
A1	1	X	0	1	X	0	0	X	X	1
A2	1	1	0	1	X	0	X	1	X	1
A3	1	1	X	1	0	0	X	1	1	1
B1	X	0	X	X	1	X	1	1	0	0
B2	1	X	1	X	1	X	1	X	0	X
B3	1	0	X	0	1	X	1	X	0	0
C1	X	X	1	1	X	0	1	1	X	0
C2	1	1	1	1	0	0	X	X	X	X
C3	X	1	X	1	0	X	1	1	1	0

Tabla 3: Observaciones eliminadas aleatoriamente (40 %)

3. Las observaciones faltantes o eliminadas en el paso anterior son reemplazadas usando métodos de imputación. Finalmente, el nuevo conjunto de datos compuesto por 9 pseudo evaluados no tiene datos faltantes y contiene vectores de respuesta diferentes a los del test original.

	1	2	3	4	5	6	7	8	9	10
A1	1	0	0	1	0	0	0	1	1	1
A2	1	1	0	1	1	0	1	1	0	1
A3	1	1	1	1	0	0	0	1	1	1
B1	1	0	1	0	1	0	1	1	0	0
B2	1	1	1	0	1	0	1	1	0	1
B3	1	0	0	0	1	0	1	1	0	0
C1	0	1	1	1	1	0	1	1	1	0
C2	1	1	1	1	0	0	1	1	1	0
C3	0	1	1	1	0	0	1	1	1	0

Tabla 4: Valores faltantes imputados

Para aplicar la técnica DuPER descrita anteriormente, se deben conocer y/o definir los siguientes aspectos:

- Número de ítems (dimensión del test, J).
- Número de evaluados (támaño de la muestra, n).
- Número de réplicas para cada evaluado (número de pseudo-evaluados).
- Tasa de eliminación para las observaciones o respuestas de los ítems.
- Método de imputación.

Así, las diferentes combinaciones entre los aspectos anteriores producen diferentes variaciones de la técnica DuPER. Según Patrick Foley (2010) se debe suponer que, si la tasa de eliminación es muy baja no se agregaría suficiente variabilidad al conjunto de datos y la técnica DuPER no sería efectiva para mejorar la estimación de los parámetros de los ítems. Por el contrario, si dicha tasa es muy alta, los vectores de respuesta imputados pueden perder su plausibilidad, resultando en vectores de respuesta irreales y malas estimaciones de los parámetros de los ítems.

3.2. Imputación múltiple con valores plausibles

La imputación múltiple es una técnica estadística diseñada para encontrar dos o más valores imputados para un valor faltante, con el fin de representar la incertidumbre que se tiene al no conocer el valor faltante (Rubin 1987). Una metodología de imputación múltiple son los valores plausibles, que matemáticamente consiste en la obtención de valores aleatorios probables o plausibles de una muestra seleccionada de la distribución condicional de la variable de interés (PISA 2012). Para mejorar dichas imputaciones se usa información auxiliar o variables de condicionamiento para cada evaluado (ICFES 2009). En el contexto de las pruebas estandarizadas, y para desarrollar matemáticamente la metodología, se puede ver la habilidad de un evaluado como un rasgo no observado.

La distribución condicional de la habilidad para cada evaluado es denotada por $p(\theta_i|\boldsymbol{X_i}, \delta, Z_i)$ en la cual se condiciona por las respuestas dadas por el evaluado a los ítems ($\boldsymbol{X_i}$) y por el vector de parámetros estimados de los mismos (δ), y la información auxiliar observable para cada evaluado Z_i. El objetivo se centra en encontrar la distribución asociada a la habilidad del i-ésimo individuo evaluado θ_i así:

$$p(\theta_i|Z_i, \boldsymbol{X_i}). \tag{17}$$

La expresión anterior se puede expresar como:

$$p(\theta_i|Z_i, \boldsymbol{X_i}) = p(\boldsymbol{X_i}|\theta_i, Z_i)p(\theta_i|Z_i). \tag{18}$$

Se deben hacer los siguientes supuestos para obtener la estimación:

1. Las respuestas de un evaluado ($\boldsymbol{X_i}$) son condicionalmente independientes de las variables auxiliares (Z_i), es decir

$$p(\boldsymbol{X_i}|\theta_i, Z_i) = p(\boldsymbol{X_i}|\theta_i). \tag{19}$$

2. Los elementos del vector $\boldsymbol{X_i}$ (las respuestas de los ítems) son condicionalmente independientes entre sí, es decir

$$p(\boldsymbol{X_i}|\theta_i) = \prod_{j=1}^{J} p(x_{ij}|\theta_i). \tag{20}$$

Del primer supuesto se puede expresar la ecuación (18) como:

$$p(\theta_i|Z_i, \boldsymbol{X_i}) = p(\boldsymbol{X_i}|\theta_i)p(\theta_i|Z_i). \tag{21}$$

De la expresión anterior se debe encontrar una distribución apropiada para $p(\theta_i|Z_i)$. Se asume que esta probabilidad tiene una distribución normal con matriz de varianzas común o constante (Σ) y con una media dada por una función lineal de las variables de condicionamiento Z_i ($Z_i\Gamma$), es decir:

$$p(\theta_i|Z_i) = \phi(\Theta_i; Z_i\Gamma, \Sigma). \tag{22}$$

en donde $\phi(\ \cdot\ ; Z_i\Gamma, \Sigma)$ denota la función de densidad multivariada normal con media $Z_i\Gamma$ y matriz de varianzas Σ. Esto sugiere ajustar el siguiente modelo de regresión:

$$\Theta_i = Z_i\Gamma + \epsilon_i, \tag{23}$$

con $\epsilon_i \sim N(0, \Sigma)$. Los parámetros a estimar son Γ y Σ y se obtienen usando el algoritmo EM.

Las principales ventajas de usar una imputación múltiple con valores plausibles radican en que, con esta metodología se puede ajustar la medida de variabilidad (error estándar) de las estimaciones de los parámetros de los ítems obtenidas, evitando así la subestimación de dicha medida y corrigiendo el sesgo (Córdoba 2016). Además, este método de imputación permite el uso de información auxiliar para realizar las estimaciones, obteniendo así más precisión.

Con esta metodología se debe realizar la estimación de los parámetros de los ítems (δ_j) de manera independiente para cada uno de los k valores plausibles y promediarlos para tener una estimación final. El número óptimo de valores plausibles es cinco, ya que, con una mayor cantidad, no se genera una ganancia significativa de información (PISA 2012). Así la estimación final de un parámetro del $j-$ésimo ítem es:

$$\hat{\delta}_j = \frac{1}{5}\sum_{k=1}^{5} \hat{\delta}_{jk}. \tag{24}$$

Los errores estándar asociados a las estimaciones que emplean valores plausibles se obtienen a partir de la raíz cuadrada de la varianza de las estimaciones, para la cual se toman en cuenta los siguientes componentes: (término 1) el promedio de la varianza de la estimación debida al modelo de TRI utilizado, y (término 2) la varianza debida a la imputación múltiple (ICFES 2009) que es un término que ajusta el error estándar para evitar su subestimación. Así, el error estándar asociado con una estimación de un parámetro del j-ésimo ítem se calcula como:

$$EE(\hat{\delta}_j) = \sqrt{\underbrace{\frac{1}{5}\sum_{k=1}^{5} var_m(\hat{\delta}_{jk})}_{\text{término 1}} + \left(1+\frac{1}{5}\right)\underbrace{\frac{1}{4}\sum_{k=1}^{5}\left(\hat{\delta}_{jk}-\hat{\delta}_j\right)^2}_{\text{término 2}}}. \tag{25}$$

En este trabajo se propone la implementación de una imputación múltiple con cinco valores plausibles para las respuestas eliminadas de los ítems específicos utilizando como información auxiliar o variables de condicionamiento las respuestas observadas de los otros ítems, por lo que cada j-ésimo ítem cuenta con un conjunto específico de predictores.

Después de este proceso se tienen $k = 5$ imputaciones (valores plausibles) para las respuestas eliminadas de los ítems, con cada una de estas imputaciones se estiman los parámetros de interés para el j-ésimo ítem ($\hat{\delta}_{jk}$) y promediando como se muestra en la ecuación (24) se obtiene el estimador final del parámetro del ítem $\hat{\delta}_j$.

3.3. Diagnóstico de recuperación de parámetros (Criterio de evaluación)

Para evaluar la eficacia de la técnica DuPER en la estimación de los parámetros de los ítems, comparando contra los parámetros obtenidos con la población completa (de ahora en adelante llamados parámetros de referencia), Patrick Foley (2010) recomienda usar el coeficiente de correlación de Pearson ρ, la raíz del error cuadrático medio (RMSE) y el sesgo, ya que estas medidas dan cuenta de la asociación, precisión y cuantifican el error entre las estimaciones y las referencias de los parámetros de los ítems. Estos criterios son usados también en Calderón & Melo (2017) y se definen como:

1. Correlación de Pearson: es una medida que expresa el grado de asociación lineal (Canavos 1988) entre los parámetros de los ítems estimados con la metodología DuPER y los parámetros de referencia, este índice se expresa por la ecuación:

$$\rho_{\hat{\delta}_j,\delta_j} = \frac{\sum_{j=1}^{J}\hat{\delta}_j\delta_j - J\bar{\hat{\delta}}\bar{\delta}}{(J-1)S_{\hat{\delta}}S_{\delta}}. \tag{26}$$

   donde J es la cantidad de ítems en el test, $\hat{\delta}_j$ y δ_j son la estimación y la referencia del parámetro de interés respectivamente para el j-ésimo ítem, $\bar{\hat{\delta}}$ y $\bar{\delta}$ son el promedio de las estimaciones y el promedio de las referencias de los parámetros de interés para todos los J ítems, $S_{\hat{\delta}}$ y S_{δ} son la desviación estándar de las estimaciones y la desviación estándar de las referencias de los parámetros de interés para todos los J ítems. Este índice varía en el intervalo [-1, 1], si $\rho = 1$ se indica que existe una relación directa perfecta entre las estimaciones y las referencias de los parámetros de los ítems, si $\rho = -1$ se indica que existe una relación inversa perfecta y si $\rho = 0$ se indica que no existe relación.

2. Raíz del error cuadrático medio (RMSE): esta medida compara las estimaciones y las referencias de los parámetros de los ítems y cuantifica la cantidad de error que existe entre ellas. Esta medida es comúnmente usada para evaluar el desempeño (precisión) de estimaciones obtenidas con modelos (T.Chai & R.Draxler 2014). Se define como:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\delta}_j - \delta_j)}{J}}. \tag{27}$$

Esta medida nunca es negativa, y si $RMSE = 0$ indica un ajuste perfecto entre las estimaciones de los parámetros de los ítems y sus respectivas referencias. Un RMSE cercano a 0 es un buen indicador de ajuste.

3. Sesgo: cuantifica la diferencia entre las estimaciones y las referencias de los parámetros de los ítems, también con esta medida es posible detectar estimaciones mayores o menores a las referencias. Calculado como:

$$\text{Sesgo} = \frac{\sum_{j=1}^{J}(\hat{\delta}_j - \delta_j)}{J}. \tag{28}$$

Es ideal que el sesgo sea 0 o muy cercano para considerar una estimación como precisa o insesgada.

4. Resultados

Se utiliza una población compuesta por 2758 individuos que fueron evaluados por 40 ítems ($J = 40$) de un módulo específico de la prueba Saber TyT aplicada en el año 2017. Con dicha población se obtienen los parámetros de los ítems usando un modelo logístico de dos parámetros (2PL) y se emplean como referencia de comparación.

En este trabajo se implementa la técnica DuPER en una muestra pequeña ($n = 250$) seleccionada aleatoriamente de dicha población y a partir de las estimaciones de los parámetros de los ítems obtenidas (con un modelo 2PL), se realiza la comparación con los parámetros de referencia. Para esto se realizan estimaciones con la combinación de diferentes aspectos de la técnica DuPER. Estas combinaciones son llamadas "escenarios" en los cuales se varía el número de réplicas para cada evaluado y la tasa de eliminación para las observaciones o ítems. Las características de los escenarios para aplicar la técnica son las siguientes:

- Número de réplicas para cada evaluado (número de pseudo evaluados = 2, 3, 4 hasta 10).
- Tasa de eliminación para las observaciones o ítems (eliminación del 10 %, 15 %, 20 % y 50 %).
- Método de imputación (Imputación múltiple con valores plausibles).

La combinación entre el número de réplicas y el número de tasas de eliminación da como resultado la cantidad de escenarios diferentes que se van a probar. De esta manera, se tienen 36 escenarios diferentes para aplicar la técnica DuPER.

Con el fin de eliminar el efecto del muestreo al seleccionar la muestra pequeña, se realizan 50 muestras aleatorias del mismo tamaño ($n = 250$), para tener como estimador final el valor esperado (promedio) de los parámetros estimados en todas las muestras. Para cada una de estas 50 muestras se aplican todos los escenarios con la técnica DuPER, es decir, que por ejemplo en el primer escenario de la primera muestra se réplica 2 veces cada evaluado, luego a cada pseudo evaluado se le elimina aleatoriamente el 10 % de las respuestas de los 40 ítems (4 ítems) y posteriormente se implementa el método de imputación múltiple con cinco valores plausibles. Este proceso es repetido con los otros 35 escenarios en la primera muestra y de la misma manera para las 49 muestras restantes. De esta manera se tienen 9000 estimaciones de los parámetros de los ítems.

Además, se realiza un análisis de datos atípicos con el método MAD (Median Absolute Deviation) el cual se puede detallar en (Leys & Klein 2013), para descartar en cada escenario las muestras con pérdida de información debida a la etapa de eliminación.

A continuación, se muestra el comportamiento de las estimaciones finales de los dos parámetros (dificultad y discriminación) del test completo en los criterios de evaluación definidos en la sección anterior, es importante resaltar que se debe tener en cuenta la escala del eje Y en las figuras, ya que visualmente esto puede generar un impacto negativo en los análisis.

En la figura 1 se evalúa la correlación de Pearson entre los parámetros estimados y los parámetros de referencia, evidenciando una correlación muy alta (mayor a 0,99) para casi todos los escenarios DuPER aplicados en ambos parámetros, los únicos escenarios en donde la correlación es menor a 0.99 es el escenario donde la tasa de eliminación es muy alta (50 %), sin embargo sigue siendo una muy buena correlación (mayor a 0.97 para la dificultad y mayor a 0.95 para la discriminación). Las mayores correlaciones para la dificultad se pueden observar con una tasa de eliminación del 20 % y muchas réplicas para cada evaluado (9 o 10). En cuanto a la discriminación, todos los escenarios (excepto tasa eliminación del 50 %) tienen correlaciones muy similares entre sí.

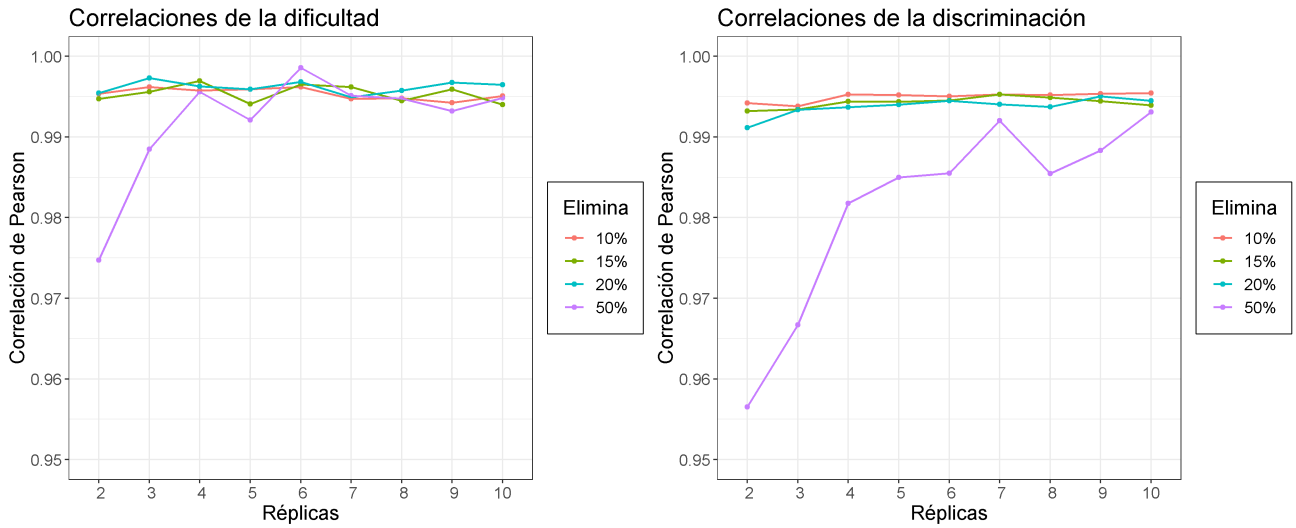

Figura 1: Correlación de Pearson entre los parámetros estimados y los de referencia.

El comportamiento de los RMSE obtenidos para las estimaciones de los dos parámetros se observa en la figura 2. Para la dificultad, los RMSE más bajos se obtienen cuando la tasa de eliminación es del 20 %, los valores obtenidos con la tasa de eliminación del 50 % son mucho más altos y variables entre las diferentes réplicas. Los valores de esta medida en las tasas de eliminación del 10 % y 15 % son mayores en la mayoría de réplicas que los observados en la tasa de eliminación del 20 %. La discriminación tiene RMSE mucho menores a los de la dificultad, para las tasas de eliminación del 10 %, 15 % y 20 % se tienen valores menores a 0.05 con diferencias despreciables entre ellas para todas las réplicas, por otro lado, la tasa de eliminación del 50 % tiene RMSE, aunque también bajos, mayores que los de las otras tasas.

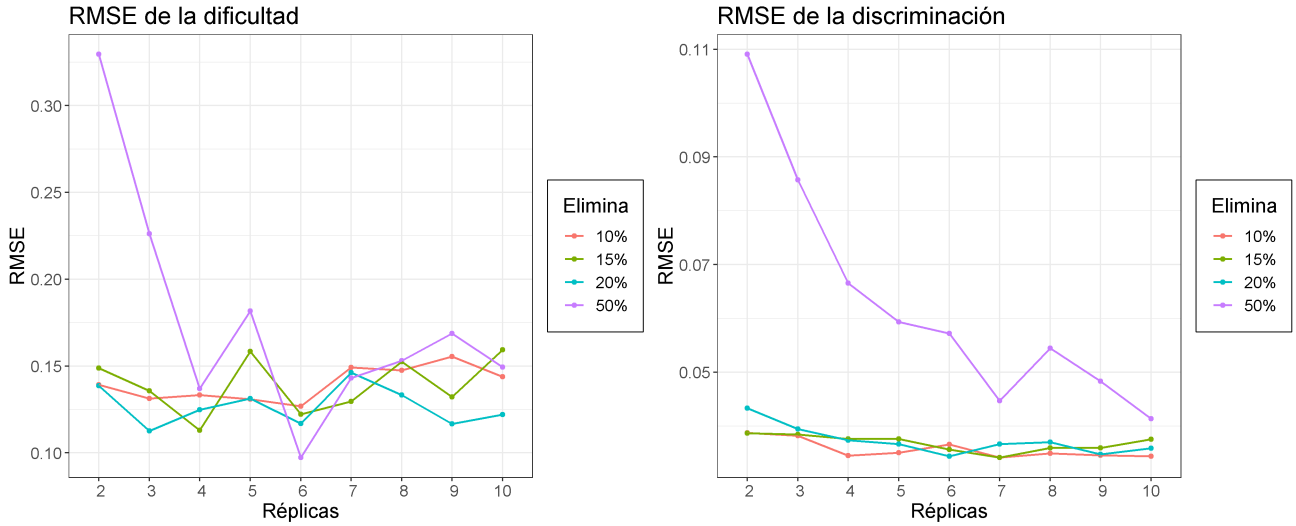

Figura 2: RMSE de los parámetros estimados.

En cuanto al sesgo de las estimaciones de los parámetros de los ítems, en la figura 3 se presentan sus valores obtenidos en la estimación de la dificultad y la discriminación con la técnica DuPER para cada uno de los escenarios. Para la dificultad de manera general en todas las tasas de eliminación y réplicas se tienen sesgos muy pequeños, casi nulos (menores a $|0.04|$). La tasa de eliminación del 10 % es la que tiene el comportamiento más constante entre las diferentes réplicas y la tasa del 50 % la que más varía entre réplicas. Para la discriminación se podría observar un sesgo positivo para las tasas de eliminación del 10 %, 15 % y 20 %, sin embargo, teniendo en cuenta la escala del sesgo en el gráfico se podría decir que este sesgo es mínimo (menores a 0.02).

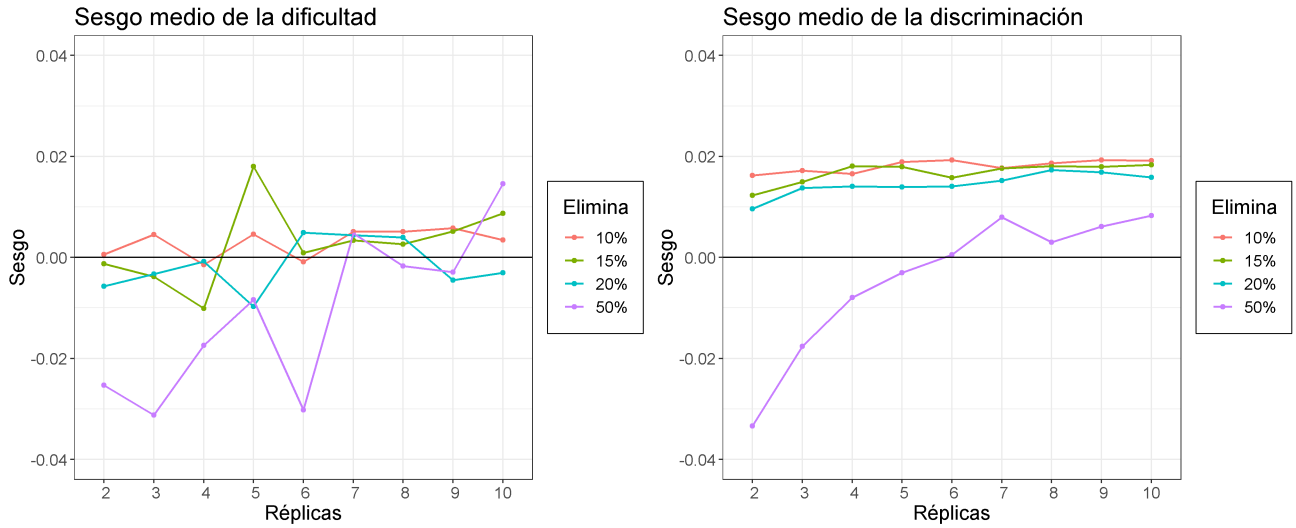

Figura 3: Sesgo de los parámetros estimados.

Conjuntamente, el comportamiento de las medidas de diagnóstico de recuperación de parámetros de las estimaciones obtenidas en cada escenario comparando con la estimación de referencia de todo el test (todos los ítems) es muy bueno, para la mayoría de escenarios se tienen coeficientes de correlación de Pearson mayores a 0.97, RMSE menores a 0.15 y sesgos alrededor de 0 para la dificultad y coeficientes

de correlación de Pearson mayores a 0.95, RMSE menores a 0.05 y sesgos positivos pero mínimos para la discriminación.

En la figura 4 se tienen las estimaciones específicas de los parámetros de un ítem, la linea negra representa el valor de referencia obtenido de la población. En ambos parámetros las estimaciones con cada escenario oscila muy cerca del valor real, además se visualiza que los escenarios con la tasa de eliminación del 50 % tienen mucha más variablilidad que los escenarios con tasas de eliminación menores.

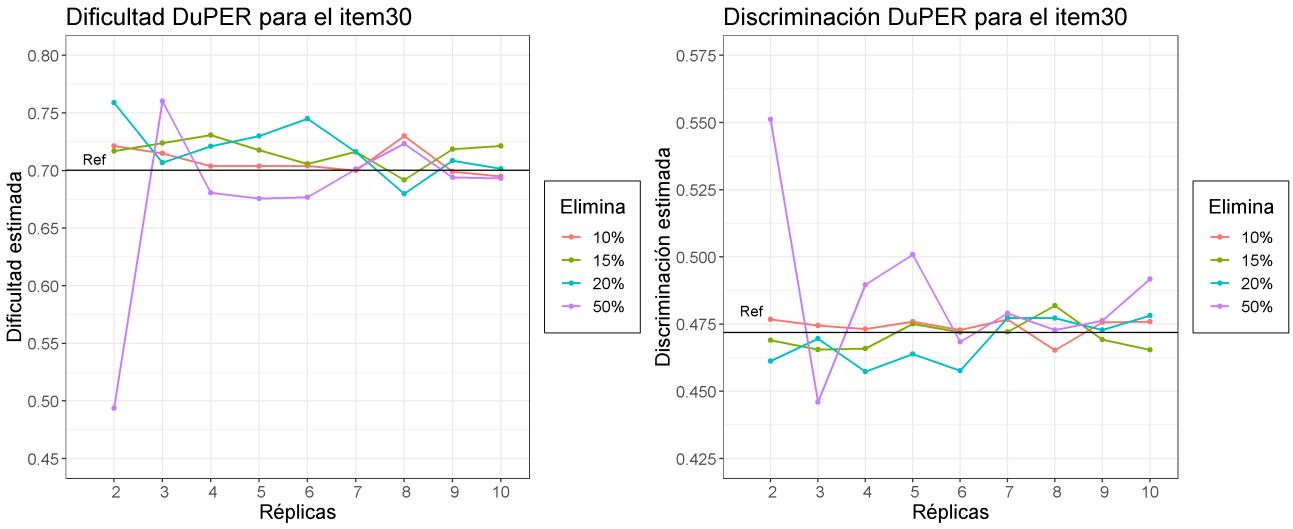

Figura 4: Parámetros estimados para un ítem

Por lo observado en las figuras del comportamiento de las medidas de diagnóstico entre el test y las estimaciones individuales de los ítems, se puede sugerir que los mejores resultados (mayor correlación, menor RMSE y sesgo nulo) se obtienen con el escenario donde la tasa de eliminación de respuestas de los ítems es del 20 % y la cantidad de réplicas por evaluado es de 8 o más. Para este escenario específico se realiza una regresión lineal entre las referencias de los parámetros de los ítems y sus respectivas estimaciones y se observa en la figura 5:

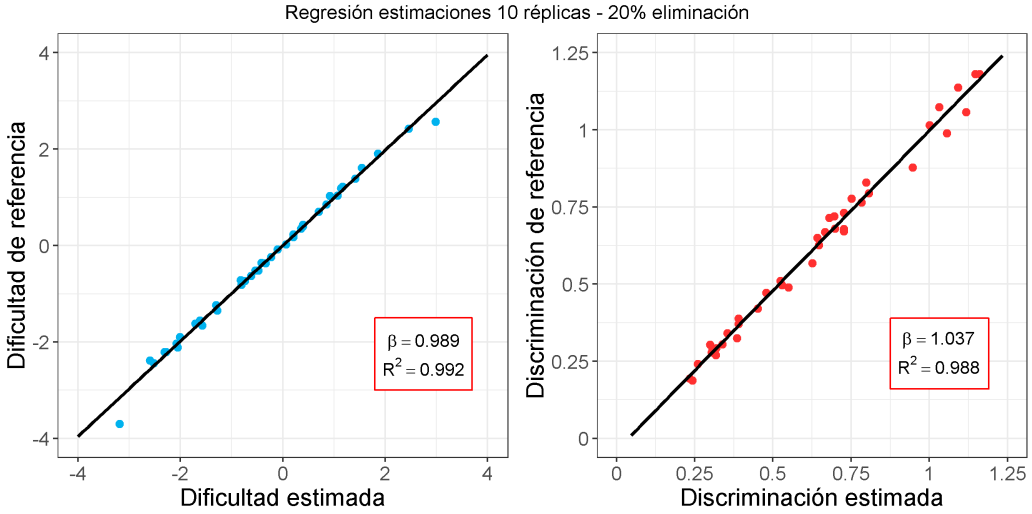

Figura 5: Regresión de los parámetros de referencia contra los estimados con 10 réplicas y 20 % de eliminación

Aquí, las líneas corresponden al modelo de regresión lineal ajustado, R^2 es la correlación ajustada por dicho modelo y β es el coeficiente de regresión de las estimaciones. Se percibe que, para valores muy bajos o muy altos en la dificultad de referencia, se tienen estimaciones con la técnica DuPER un poco sesgadas, generando diferencias en este parámetro. Sin embargo, cuando las dificultades de referencia está en un rango medio, las estimaciones obtenidas con la técnica DuPER son muy precisas, generando diferencias nulas. La relación entre estimaciones y referencias se cuantifica con el coeficiente $\beta = 0.989$ el cual, por su cercanía a 1 indica una relación casi perfecta entre las estimaciones de los parámetros obtenidas y las estimaciones de referencia. En cuanto a la discriminación, se considera que no hay diferencias significativas entre las estimaciones y las referencias teniendo en cuenta la escala de la figura, $\beta = 1.037$ indica una muy buena relación entre los parámetros comparados.

Adicionalmente, para mostrar el ajuste que el método de imputación múltiple con valores plausibles hace al error estándar de las estimaciones, en el escenario escogido (10 réplicas y 20 % eliminación) se utiliza la técnica DuPER con el algoritmo EM para imputar (una sola vez) las respuestas de los ítems eliminadas y se estiman los parámetros de los ítems y sus respectivos errores estándar. En la tabla 5 se muestra el error estándar de las estimaciones de los parámetros para dos ítems obtenidos con la población completa (referencia), con la técnica DuPER usando imputación múltiple con valores plausibles y con la imputación usando el algoritmo EM.

Estimación del	**Ítem25**		**Ítem35**	
Error Estándar	**Dificultad**	**Discrim**	**Dificultad**	**Discrim**
Referencia	0.124	0.047	0.131	0.048
Valores plausibles	0.125	0.058	0.134	0.058
Algoritmo EM	0.015	0.002	0.014	0.003

Tabla 5: Comparación de la estimación del error estándar

Como se mencionó antes, una de las principales ventajas de usar una imputación múltiple con valores plausibles radica en que este método de imputación se ajusta el error estándar evitando la su subestimación y corrigiendo el sesgo. Se puede notar que con el uso de valores plausibles aproxima mucho la estimación del error estándar a los valores de referencia. Por el contrario, hacer una sola imputación subestima en gran medida el error estándar.

5. Conclusiones

A partir de la implementación de la técnica DuPER en diferentes escenarios con la propuesta de imputación múltiple con valores plausibles en una muestra pequeña (insuficiente para garantizar estimaciones de parámetros de los ítems precisas), se puede concluir que el comportamiento de la precisión de las estimaciones de los parámetros entre las diferentes réplicas de los escenarios no varía mucho, en cambio se evidencia que la tasa de eliminación si presenta un impacto más directo e importante en dichas estimaciones, por ejemplo se evidenció que una tasa de eliminación del 50 % induciría mucha variabilidad en las estimaciones. Con una tasa de eliminación moderada en cada vector de respuesta (20 %) y una cantidad grande de réplicas (8 o más) se obtienen las mejores estimaciones de los parámetros de los ítems, con correlaciones mayores a 0.995, RMSE menores a 0.15 y sesgo nulo.

Además, la propuesta de usar imputación múltiple con valores plausibles permite utilizar información auxiliar para realizar una mejor estimación y agregar un término de error debido a la imputación permitiendo corregir o ajustar el error estándar de las estimaciones de los parámetros de los ítems, y así disminuyendo la probabilidad de cometer error tipo I.

6. Discusión

Según los hallazgos hechos por Patrick Foley (2010), la técnica DuPER aplicada en una muestra pequeña no es efectiva para producir estimaciones precisas de los parámetros de los ítems, pues estas tienen altos RMSE y bajas correlaciones con las referencias. Sin embargo, estas conclusiones se obtuvieron a partir de escenarios netamente simulados por lo que pueden estar sesgadas por las condiciones y características de simulación. En este trabajo se usó información real de la prueba Saber TyT en donde se encontró que la aplicación de la técnica DuPER en una muestra pequeña, produce estimaciones precisas de los parámetros de los ítems permitiendo así el uso de modelos de TRI sobre muestras pequeñas. Esto implica que a pesar de que no exista una población suficientemente grande o que no se cuente con suficientes recursos logísticos o económicos para garantizar una cantidad mínima de evaluados, se pueden utilizar modelos de TRI para tener estimaciones precisas tanto de los parámetros de los ítems como de las habilidades de los evaluados.

No obstante, para que la técnica DuPER sea efectiva, la muestra de individuos usada debe ser representativa de la población que se quiere evaluar y el comportamiento de los ítems debe ser óptimo en el sentido de que el test no tenga confiabilidad baja o sospechas de multidimensionalidad. Si alguna de estas condiciones no se cumple, la fluctuación en las estimaciones de los parámetros de los ítems puede ser considerable aumentando así el sesgo.

Para reducir costos significativamente en los pilotajes de los test, bastaría con garantizar la selección de una muestra representativa de la población que se desea evaluar y con la aplicación de la técnica DuPER generar la cantidad de evaluados requerida para usar modelos de TRI.

Para trabajos futuros: en la etapa de eliminación es posible que se dañen los vectores o strings de respuesta produciendo estimaciones sesgadas, esto pasa cuando los ítems tienen muchas respuestas correctas o muchas respuestas incorrectas por lo que es fácil perder información con la eliminación de observaciones. Es posible generalizar el tratamiento de los vectores de respuesta después de la eliminación implementando un criterio estadístico que determine la información I del vector resultante con el interés de establecer una regla de decisión que juzgue si la perdida de información es tal que las estimaciones obtenidas son poco probables y por lo tanto sesgadas; así con esta regla se puede tomar la decisión de descartar o conservar el vector, es decir, si $I(\boldsymbol{X_i}) > \alpha$ se conserva el vector $\boldsymbol{X}$ del $i-$ésimo individuo, en otro caso se descarta y se seleccionan de nuevo aleatoriamente las observaciones a eliminar y se repite el proceso de validación.

También es posible establecer una cantidad diferente de réplicas por individuos o grupos de individuos desde una metodología de TRI, usando un modelo nominal para establecer una métrica del conjunto de datos observados y así poder hacer una selección objetiva y sistemática de habilidades y de esta manera establecer matemáticamente un criterio para realizar la replicación.

Referencias

Calderón, C. & Melo, O. (2017), 'Un estudio de simulación para la evaluación de diseños y tamaños muestrales requeridos en la estimación de parámetros de un modelo politómico de teoría de respuesta al ítem y parámetros poblacionales de interés.', *Universidad Nacional de Colombia* .

Canavos, G. C. (1988), 'Probabilidad y estadística, aplicaciones y métodos', *McGRAW HILL* .

Córdoba, M. F. (2016), 'Una aplicación de valores plausibles a la calificación de pruebas estandarizadas vía simulación', *Comunicaciones en Estadística - Universidad Santo Tomás* **9**(1).

de Ayala, R. J. (2009), 'The theory and practice of item response theory', *THE GUILFORD PRESS - New York London* **1**(1).

Dempster, A. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM alforithm', *Journal of the Royal Statistical Society* **39**(1).

Fox, J. P. (2010), 'Bayesian item response modeling', *Statistics for Social and Behavioral Sciences* **1**(1).

Hambleton, R. K. (1989), 'Principles and selected applications of item response theory', *R. L. Linn, Educational measurement* **3**(1).

ICFES (2009), Informe técnico saber 5o. y 9o. 2009, Technical report, Instituto Colombiano para la Evaluación de la Educación.

Leys, C. & Klein, O. (2013), 'Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median', *Journal of Experimental Social Psychology - ELSEVIER* .

Linacre, J. M. (1994), 'Sample size and item calibration stability', *Rasch Measurement Transactions* **7**(4).

Patrick Foley, P. (2010), 'Improving IRT parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique', *Educational Assessment, Evaluation, and Research Commons, Educational Psychology Commons, and the Quantitative Psychology Commons* **7**(1).

PISA (2012), Pisa 2012 technical report, Technical report, Organisation for Economic Co-operation and Development (OECD).

Reckase, M. D. (2009), 'Multidimensional item response theory', *Statistics for Social and Behavioral Sciences* **1**(1).

Rubin, D. B. (1987), 'Multiple imputation for nonresponse in surveys', *New York: Wiley* .

Sahin, A. & Anil, D. (2016), 'The effects of test length and sample size on item parameters in item response theory', *KURAM VE UYGULAMADA EGITIM BILIMLERI EDUCATIONAL SCIENCES: THEORY AND PRACTICE* **1**(1).

T.Chai & R.Draxler (2014), 'Root mean square error (rmse) or mean absolute error (mae)', *Geoscientific Model Development* .