Polo del Conocimiento

Pol. Con. (Edición núm. 69) Vol. 7, No 4
Abril 2022, pp. 556-575
ISSN: 2550 - 682X
DOI: 10.23857/pc.v7i4.3840

Metodología de equiparación en evaluaciones estandarizadas mediante un algoritmo estadístico

Equating methodology in standardized assessments using a statistical algorithm

Metodologia de correspondência em avaliações normalizadas utilizando um algoritmo estatístico

Héctor Salomón Mullo-Guaminga [I]
hmullo@espoch.edu.ec
https://orcid.org/0000-0001-8448-4652

Jessica Alexandra Marcatoma-Tixi [II]
jessica.marcatoma@unach.edu.ec
https://orcid.org/0000-0001-9531-3234

Washington Javier Carrasco-Tuston [III]
washington.carrasco@iess.gob.ec
https://orcid.org/0000-0002-6544-8960

Oswaldo Villacrés-Cáceres [IV]
ovillacres@espoch.edu.ec
https://orcid.org/0000-0002-5894-5248

Correspondencia: hmullo@espoch.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de Investigación

***Recibido:** 27 de febrero de 2022 ***Aceptado:** 24 de marzo de 2022 * **Publicado:** 01 abril de 2022

I. Ingeniero en Estadística Informática, Máster Universitario en Estadística Aplicada, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
II. Ingeniero en Estadística Informática, Máster Universitario en Estadística Aplicada, Universidad Nacional de Chimborazo, Riobamba, Ecuador.
III. Biofísico, Máster en Sistemas de Gestión Ambiental, Instituto Ecuatoriano de Seguridad Social, Cuenca, Ecuador.
IV. Ingeniero en Sistemas Informáticos, Magíster en Gestión de Bases de Datos, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.

Resumen

Este trabajo estudia la forma de comparar resultados de los estudiantes en pruebas estandarizadas a gran escala, al respecto plantea el desarrollo de un algoritmo en software estadístico R de una metodología de 5 pasos para el desarrollo de una equiparación entre dos formas de prueba (X e Y) en el contexto de diseños de equiparación, donde los grupos son no equivalentes, con covariables e ítems ancla. El resultado principal muestra que, para el desarrollo con éxito de la equiparación se debe realizar: i) Calibración de la forma de prueba X e Y con covariables mediante regresión latente; ii) Calibración concurrente de las formas de prueba X e Y; iii) Calibración de la forma de prueba X con parámetros de los ítems fijos obtenidos del paso dos, donde los ítems fijos son aquellos ítems considerados como anclas; iv) Obtención de constantes de transformación lineal, utilizando el método de momentos y el enfoque de valores plausibles y, v) Escalamiento de la forma de prueba Y a la escala de X utilizando los resultados del paso anterior. Este resultado muestra que es factible el desarrollo de una equiparación en el contexto planteado y que cuando se pretenda compara las habilidades de los estudiantes que rindieron dos formas de prueba diferentes, siempre se debe realizar una equiparación para que sean comparables las diferencias en la media y la desviación estándar de las distribuciones de habilidad estimadas de las diferentes formas de prueba, tomado a una de ellas como escala base.

Palabras Claves: Evaluación estandarizada; Valor Plausible; Equiparación; Psicometría.

Abstract

This work studies the way to compare student results in large-scale standardized tests, in this regard it proposes the development of an algorithm in statistical software R of a 5-step methodology for the development of an equating between two test forms (X and Y) in the context of matching designs, where the groups are non-equivalent, with covariates and anchor items. The main result shows that, for the successful development of the matching, the following must be carried out: i) Calibration of the test form X and Y with covariates by means of latent regression; ii) Concurrent calibration of test forms X and Y; iii) Calibration of the test form X with parameters of the fixed items obtained from step two, where the fixed items are those items considered as anchors; iv) Obtaining linear transformation constants, using the method of moments and the plausible values approach and v) Scaling the test form Y to the scale of X using the results of the previous step. This result shows that the development of an equalization in the matched context is feasible and

that when it is intended to compare the abilities of the students who took two different forms of test, an equalization must always be carried out so that the differences in the mean and the standard deviation of the estimated ability distributions of the different test forms, taking one of them as the base scale

Keywords: Standardized evaluation; Plausible Value; Equating; Psychometry.

Resumo

Este artigo estuda como comparar as pontuações dos estudantes em testes padronizados de grande escala, desenvolvendo um algoritmo em software estatístico R de uma metodologia de 5 etapas para o desenvolvimento de uma correspondência entre duas formas de teste (X e Y) no contexto de desenhos de correspondência, onde os grupos são não equivalentes, com covariáveis e itens de âncora. O resultado principal mostra que, para o desenvolvimento bem sucedido da correspondência, deve ser realizado o seguinte: (i) Calibração do formulário de ensaio X e Y com covariáveis usando regressão latente; (ii) Calibração simultânea dos formulários de ensaio X e Y; (iii) Calibração do formulário de ensaio X com parâmetros de item fixo obtidos do passo dois, em que os itens fixos são os itens considerados como âncoras; (iv) Obtenção de constantes de transformação linear, usando o método dos momentos e a abordagem do valor plausível e, (v) Escala do formulário de ensaio Y à escala de X usando os resultados do passo anterior. Este resultado mostra que é viável desenvolver a equação no contexto dado e que ao tentar comparar as capacidades dos estudantes que fizeram duas formas de teste diferentes, a equação deve ser sempre feita de modo a que as diferenças na média e desvio padrão das distribuições de capacidades estimadas das diferentes formas de teste sejam comparáveis, tomando uma delas como a escala de base.

Palavras-chave: Avaliação normalizada; Valor plausível; Equação; Psicometria.

Introducción

Las pruebas estandarizadas en el sistema educativo del Ecuador han adquirido un papel importante en los últimos años, principalmente para el acceso a la educación superior. Cada vez más, las vidas de los estudiantes se ven influenciadas por las pruebas estandarizadas, ya que un impulso social por la responsabilidad educativa ha llevado a un aumento dramático en el uso de estas evaluaciones

a nivel nacional. El propósito de las pruebas estandarizadas es hacer que un gran número de sustentantes desarrollen una sola prueba, luego comparar cualquier puntaje individual con todos los demás para ver su posición relativa. Los resultados se publican en una curva de campana que indica dónde se encuentra un puntaje dentro de los estándares estadísticos. Las pruebas estandarizadas se administran a grandes grupos que suman al menos miles, a veces millones. Para que los resultados sean lo más válidos posible, y así "estandarizar" la administración de la evaluación, las pruebas son: i) desarrolladas en el mismo instante por todos los sustentantes, ii) con administradores de la prueba consistentes, iii) con el mismo tiempo máximo de resolución de la prueba para todos los sustentantes y iv) con instrumentos de evaluación equivalentes entre sustentantes. Por lo tanto, "estandarizado" significaba que la prueba es estándar o lo mismo en tres aspectos: i) ítems/equivalentes, ii) instrucciones y iii) asignación de tiempo (Burke, 1999).

En general en las pruebas estandarizadas los aspectos de las instrucciones y de asignación de tiempo están bien controladas y son relativamente fáciles de estandarizar. En contraste, el aspecto de ítems/equivalentes es una cuestión importante y medular, debido a que normalmente se aplican a los sustentantes varias formas de prueba[1], que no se puede asegurar que sean equivalentes en términos de dificultad. Por lo tanto, no es correcto comparar directamente los puntajes entre sustentantes que rindieron la evaluación en dos formas de prueba diferentes, más bien se debe realizar un proceso de equiparación de puntajes entre estas dos formas de prueba.

Por otro lado, la teoría de la respuesta al ítem (TRI) es una de las teorías psicométricas más utilizadas en la actualidad que sirven para modelar las respuestas a los ítems de los sustantivos y se denomina proceso de calibración. El desarrollo inicial de los modelos TRI tuvo lugar en la segunda mitad del siglo XX. Primero, Rasch (1960) desarrolló un modelo para analizar datos categóricos. A continuación, Lord y Novick (1968) escribieron capítulos sobre la teoría de la estimación de rasgos latentes, que dio origen a una nueva forma de análisis de datos en las pruebas. Los métodos TRI se utilizan hoy en día en el desarrollo de pruebas, bancos de ítems, análisis de datos, análisis del funcionamiento diferencial de ítems, pruebas adaptativas y equiparación de formas de prueba (Kolen y Brennan, 2014).

La equiparación es un proceso estadístico que se utiliza para ajustar los puntajes en las diferentes formas de prueba de manera que estos puntajes se puedan usar indistintamente de la forma (Kolen y Brennan, 2014). Después de la equiparación, las formas de prueba producen puntuaciones

[1] Una forma de prueba es una acomodación específica de las preguntas en una evaluación.

escaladas que se pueden usar indistintamente, aunque se basen en diferentes conjuntos de ítems. Es importante señalar que el ajuste estadístico por equiparación se realiza cuando el contenido de las formas de prueba es el mismo, es decir, se evalúa el mismo constructo.

Ahora, para realizar una equiparación entre dos formas de prueba es necesario datos de puntuación de los sustentantes que permita diferenciar en la distribución de puntajes, la dificultad de las formas de prueba y las habilidades de los sustentantes. El modo de realizar esta diferenciación es mediante la utilización de personas o ítems comunes en las formas de prueba. Existen varios métodos para realizar este trabajo y se conocen como diseños de equiparación. Dentro de los diseños más utilizados se tiene dos: i) Diseño de grupos no equivalentes con prueba de anclaje (por sus siglas en inglés NEAT: Nonequivalent groups with anchor test), es decir, una evaluación donde existen dos grupos de sustentantes, sin estudiantes en común, y dos formas de prueba que tiene un grupo de ítems en común, llamado prueba de anclaje, y además asigna cada forma de prueba a un grupo distinto. ii) diseño de grupos no equivalentes con covariables (por sus siglas en inglés NEC: Nonequivalent groups with covariates) que es similar al diseño NEAT sin una prueba de anclaje y con covariables producto de una encuesta de factores asociados del rendimiento de los sustentantes que se aplica previo a la evaluación formal. Si bien es cierto existen buenos métodos para desarrollar una equiparación mediante los diseños NEAT y NEC. Sin embargo, existen escasos trabajos que describen el desarrollo de una equiparación de dos formas de prueba en el contexto de un diseño con una prueba de anclaje y covariables, es decir, con un diseño combinado de NEAT y NEC. En el mismo sentido, de nuestro conocimiento no existe un trabajo que describa el desarrollo práctico de este tipo de equiparación.

En consecuencia, el presente trabajo presenta una metodología para desarrollar en la práctica una equiparación entre dos formas de prueba utilizando TRI en el contexto de una combinación de los diseños NEAT y NEC. Es importante, indicar que este diseño combinado hace uso de una forma generalizada del modelo de Rasch, dado en Adams, Wilson y Wu (1997) y Adams y Wu (2007). Entonces, con este objetivo se describe a continuación siguiendo a Gonzáles y Wiberg (2017) los modelos TRI, equiparación de las puntuaciones de TRI mediante calibraciones separadas, calibración concurrente y calibraciones de parámetros de ítems fijos.

Equiparación utilizando TRI en el contexto del diseño NEAT

Modelos TRI

Sea X_{ij} la variable aleatoria que denota la respuesta del individuo i en el ítem j en la forma de prueba X (la notación y las definiciones que siguen se pueden adaptar fácilmente para la forma de prueba Y). Suponiendo $i = 1, \dots, n_x$ sustentantes y $j = 1, \dots, J_x$ ítems, los datos observados se pueden acomodar en una matriz $n_x \times J_x$ donde cada fila contiene el patrón de respuesta de cada sustentante. Tenga en cuenta que en el caso de que los elementos se califiquen de forma binaria (es decir, 1 si la respuesta es correcta y 0 en caso contrario), las puntuaciones de suma $X_i = \sum_{j=1}^{J_x} X_{ij}$ se pueden calcular y utilizar como la puntuación del sustentante i. Una forma alternativa de producir las puntuaciones de los examinados es utilizar modelos de TRI. Los modelos de TRI para ítems con puntuación binaria especifican la probabilidad del sustentante de una respuesta correcta en un ítem de prueba basado tanto en el parámetro de habilidad de un sustentante, θ_i, como en un vector de características del ítem, w_j (por ejemplo, el nivel de dificultad del ítem, la discriminación del ítem, etc.). Cuando se supone que $\theta_i \sim N(0, \sigma_\theta^2)$ el modelo estadístico se convierte en

$$(X_{ij}|\theta_i, w_j) \sim Bernoulli\left(\pi(\theta_i, w_j)\right)$$

Donde $\pi(.)$ se conoce como la curva característica del ítem (ICC). En particular, el modelo TRI logístico de tres parámetros (3PL) emplea $\pi(\theta_i, w_j) = c_j + (1 - c_j)\Psi\left(Da_j(\theta_i - b_j)\right)$, donde $w_j = (a_j, b_j, c_j)$, D es una constante de escala, y $\Psi(x) = exp(x)/[1 + exp(x)]$ es la función logística estándar. Los componentes a_j, b_j, c_j en w_j son los parámetros de discriminación, dificultad y pseudo adivinación del ítem, respectivamente. Bajo esta especificación, tenemos

$$\pi_{ij} = Pr\ Pr\left(X_{ij} = 1|\theta_i, w_j\right) = c_j + (1 - c_j)\frac{exp\left[Da_j(\theta_i - b_j)\right]}{1 + exp\left[Da_j(\theta_i - b_j)\right]}$$

Otros modelos TRI pueden verse como casos especiales del modelo 3PL. Por ejemplo, el modelo TRI logístico de dos parámetros (2PL) se obtiene configurando $c_j = 0$ para todo j, mientras que el modelo TRI logístico de un parámetro (1PL) establece adicionalmente que a_j sea igual a 1.

Equiparación de las puntuaciones de TRI mediante calibraciones separadas

Para equiparar los puntajes de TRI, necesitamos una función de transformación, φ, que mapee la escala de TRI en la forma de prueba X a la forma de prueba Y, es decir, $\varphi : \Theta_X \mapsto \Theta_Y$, donde Θ es el rango de las puntuaciones de TRI y se denomina escala TRI. A menudo se asume una escala normal estándar, por lo tanto $\Theta \approx [-3, 3]$.

Linking de parámetros

Debido a que los parámetros de diferentes formas de prueba deben estar en la misma escala, el linking de parámetros de TRI se realiza para colocar las estimaciones de los parámetros de TRI de **calibraciones separadas** de dos formas de prueba, en una escala común. Esto es necesario en particular cuando se realiza la equiparación bajo el diseño NEAT. La transformación $\varphi : \Theta_X \mapsto \Theta_Y$ se ha asumido típicamente que es una ecuación lineal utilizada para convertir las puntuaciones de TRI como $\theta_{Yi} = A\theta_{Xi} + B$.

Las relaciones entre los parámetros de los ítems en las dos formas de prueba son las siguientes:

$$a_{Yj} = \frac{a_{Xj}}{A}$$

$$b_{Yj} = Ab_{Xj} + B$$

$$c_{Yj} = c_{Xj},$$

Donde A y B son constantes de linking (también conocidas como coeficientes de equiparación) que deben estimarse. Los índices X e Y se utilizan para diferenciar entre las escalas. Es importante indicar que el parámetro de pseudo adivinación es independiente de la escala.

Métodos de momentos para estimar coeficientes de equiparación

Estos métodos utilizan diferentes momentos, es decir, las medias (momento 1) y las desviaciones estándar (momento 2) de las estimaciones de los parámetros de los ítems comunes, para obtener los coeficientes A y B de equiparación. En los métodos de mean-mean y mean-sigma, las medias y las desviaciones estándar se definen solo en el conjunto de ítems comunes (conocidas como ítems ancla) entre las formas de prueba X e Y. Sean μ_{aX} y μ_{aY} la media de las estimaciones de los parámetros de discriminación de los ítems tomadas solo en el conjunto de ítems comunes, y sean σ_{aX} y σ_{aX} las correspondientes desviaciones estándar.

El método de mean-mean define el coeficiente de igualación A cómo $A = \frac{\mu_{aX}}{\mu_{aY}}$, y el método mean-sigma define la constante A como $A = \frac{\sigma_{bX}}{\sigma_{bY}}$.

Aunque se utilizan diferentes A's en estos tres métodos, la constante B se define en cada uno de estos casos como $B = \mu_{b_Y} - A\mu_{b_X}$.

Equiparación de las puntuaciones de TRI mediante Calibración Concurrente

Anteriormente, las puntuaciones de TRI se vincularon mediante una función de transformación lineal $\varphi : \Theta_X \mapsto \Theta_Y$. Sin embargo, existen métodos alternativos que no necesitan una transformación de equiparación para realizar la equiparación y, en cambio, los parámetros se vinculan en una escala común durante la rutina de estimación. Este es el caso de la Calibración Concurrente y Calibración de Parámetros de Ítems Fijos. El primero lo describimos en esta sección, mientras que el segundo en la siguiente sección.

En el método de calibración concurrente Wingersky y Lord (1984), los parámetros obtenidos de los datos de ambas formas de prueba (es decir, X en Y) se estiman juntos en una sola ejecución del software TRI donde se asumen distribuciones de habilidad separadas en las dos poblaciones. Los ítems que no son comunes se tratan como no alcanzados por el programa. En la Tabla siguiente se muestra una representación esquemática de la estructura de datos necesaria para realizar una calibración concurrente. Los ítems no alcanzados por la calibración concurrente están representados por NA en la Tabla 1, mientras que los alcanzados se representan con un asterisco de color azul.

Tabla 1. Estructura de datos en la calibración concurrente, para dos formas

		X		***A***		***Y***	
	Id / Id ítem	ítem 1	Ítem 2	ítem 1	Ítem 2	ítem 1	Ítem 2
P	1	*	*	*	*	NA	NA
	2	*	*	*	*	NA	NA
Q	1	NA	NA	*	*	*	*
	2	NA	NA	*	*	*	*
	3	NA	NA	*	*	*	*

Fuente: González y Wiberg, 2017

Como se mencionó anteriormente, debido a que las dos poblaciones de examinados podrían diferir significativamente en habilidad, el software TRI utilizado para la calibración simultánea debe tener una función que admita múltiples grupos para la estimación. Esto significa que se asume una distribución diferente de habilidades para cada uno de los grupos en la calibración. Es posible que los parámetros de dificultad de los ítems sean sobreestimados para los ítems de la forma de prueba menos difícil y subestimados para los ítems de la forma de prueba más difícil si no se tienen en cuenta las diferencias de habilidad de los grupos (DeaMars, 2002).

Equiparación de las puntuaciones de TRI mediante Calibración de parámetros de ítems fijos

En el método de calibración de parámetros de ítems fijos Kim (2006), los valores estimados para ítems comunes de calibraciones anteriores se fijan en la calibración de otras formas de prueba. Suponiendo que se administra primero la forma de prueba Y, las estimaciones de los parámetros para la parte de ítems comunes se utilizan y se tratan como fijas al calibrar la forma de prueba X. De esta manera, la escala X se ve obligada a estar en la escala Y.

Calibración de la forma X e Y con covariables mediante regresión latente

La regresión latente de Adams y Wu (2007) y Adams, Wilson y Wu (1997), se muestra a continuación. En primer lugar, es importante observar que la descripción de un modelo estructural de respuesta al ítem requiere la especificación de dos componentes: un modelo de respuesta al ítem condicional $f_{\mathbf{x}}(\mathbf{x};\boldsymbol{w}|\theta)$ y un modelo poblacional $f_{\theta}(\theta;\boldsymbol{\alpha})$, donde $\mathbf{x}$ es un vector de observaciones sobre el ítems, $\boldsymbol{w}$ es un vector de parámetros que describen esos ítems, θ es una variable aleatoria latente (típicamente habilidad), y $\boldsymbol{\alpha}$ simboliza un conjunto de parámetros que caracterizan la distribución de θ. El modelo de población describe la variación entre estudiantes en el rasgo latente de interés, y el modelo de respuesta condicional al ítem describe la probabilidad de observar un conjunto de respuestas de los ítems condicionadas al nivel de un individuo en el rasgo latente de interés. En el caso de que la población se considere normal, $\boldsymbol{\alpha} = (\mu, \sigma^2)$. Este es un modelo estructural porque θ es una variable aleatoria, es decir, no tiene un valor desconocido fijo.

Modelo poblacional

El modelo de respuesta al ítem es un modelo condicional, en el sentido de que describe el proceso de generar respuestas al ítem condicionadas a la variable latente, $\boldsymbol{\theta}$. La definición completa del modelo, por lo tanto, requiere la especificación de una densidad, $f_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\alpha})$, para la variable latente, $\boldsymbol{\theta}$. La práctica más común para especificar modelos de respuesta de ítems marginales

Pol. Con. (Edición núm. 69) Vol. 7, No 4, Abril 2022, pp. 556-575, ISSN: 2550 - 682X

unidimensionales es asumir que los estudiantes han sido muestreados de una población normal con media μ y varianza σ^2. Eso es,

$f_\theta(\theta;\alpha) \equiv f_\theta(\theta;\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\, exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)$, o equivalentemente $\theta = \mu + E$, donde $E \sim N(0,\sigma^2)$.

La ecuación $\theta = \mu + E$ es un modelo lineal muy simple para la variación entre estudiantes en θ. Una extensión natural de esta ecuación es reemplazar la media, μ, con el modelo de regresión $\boldsymbol{Y}_n^T\beta$, donde $\boldsymbol{Y}_n$ es un vector de μ, de valores fijos y conocidos para el estudiante n, y β es el vector correspondiente de coeficientes de regresión. Por ejemplo, $\boldsymbol{Y}_n$ podría estar constituido por variables de estudiantes como género o nivel socioeconómico, etc. Entonces el modelo de población se convierte en $\boldsymbol{\theta}_n = \boldsymbol{Y}_n^T\beta + E_n$, donde asumimos que las E_n son independientes e idénticamente distribuidas mediante una norma de media cero y varianza σ^2 de modo que

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\boldsymbol{\alpha}) = f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\beta,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\, exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\theta}_n - \boldsymbol{Y}_n^T\beta)^T(\boldsymbol{\theta}_n - \boldsymbol{Y}_n^T\beta)\right),$$

Una normal con media $\boldsymbol{Y}_n^T\beta$ y varianza σ^2.

Entre las ventajas de utilizar covariables, $\boldsymbol{Y}_n$, en la estimación de los parámetros de los ítems y la habilidad de los sustentantes tenemos que puede conducir a una mayor precisión en la estimación de los parámetros de los ítems y de la habilidad de los sustentantes. Es importante mencionar que según Mislevy y Sheehan (1989) para asegurar estimaciones consistentes de los parámetros de los ítems, la información colateral ($\boldsymbol{Y}_n$) debe usarse en la estimación donde se utilizó en la selección de ítems. Por ejemplo, cuando a los estudiantes de diferentes niveles de grado se les dan diferentes formas de prueba, que se ajustan a sus habilidades esperadas a través de la información de edad o grado, la estimación del parámetro de máxima probabilidad marginal (MML: marginal maximum likelihood) posterior será inconsistente a menos que la información de edad o nivel de grado se utilice como una variable colateral.

Modelo de respuesta al ítem condicional

La regresión del vector de respuesta sobre los parámetros del ítem y del sustentante es

$f(\mathrm{x}; w|\theta) = \Gamma(\theta, w)\, exp\big(\mathrm{x}^T(\boldsymbol{M}\theta + \boldsymbol{H}w)\big)$, con $\Gamma(\theta, w) = \big(\sum_{\boldsymbol{z}\in\Omega} exp\big(\boldsymbol{z}^T(\boldsymbol{M}\theta + \boldsymbol{H}w)\big)\big)^{-1}$, donde

- Ω es el conjunto de todos los posibles vectores de respuesta.
- $\boldsymbol{H}$ es una matriz de diseño igual a $\big(a_{11}, a_{12}, \dots, a_{1k_1}, a_{21}, \dots, a_{2k_2}, \dots, a_{Ik_I}\big)^T$, en el cual, $i = 1, \dots, I$; $k = 1, \dots, K_i$; I es la cantidad de ítems y $K_i + 1$ es el número de alternativas de respuesta al ítem i.
- $\boldsymbol{M}$ es una matriz de puntuación (valoración o calificación) igual a $(M_1^T, M_2^T, \dots, M_I^T)$ con $M_i = \big(m_{i1}, m_{i2}, \dots, m_{ik_I}\big)^T$; $m_{ik} = (m_{ik1}, m_{ik2}, \dots, m_{ikD})^T$ y m_{ikd} es el nivel de puntuación que se asigna a cada tipo de respuesta observada en la categoría k, ítem i de la dimensión d. La forma multidimensional del modelo asume que un conjunto de rasgos D subyacen a las respuestas de los sustentantes. Los rasgos latentes D definen un espacio latente D-dimensional. El vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_D)^T$ representa la posición de un individuo en el espacio latente D-dimensional.

Metodología

Participantes: Los participantes de una evaluación educativa estandarizada a gran escala en el Ecuador son principalmente personas que desean ingresar a la educación superior de grado y/o posgrado, para seleccionar a nuevos profesores del magisterio, para procesos de homologación o reingreso al Sistema Educativo Nacional, para evaluar la gestión educativa y habilidades de razonamiento de aspirantes a cargos directivos.

Proceso: Los datos de evaluaciones estandarizadas a gran escala en el Ecuador se obtienen a partir de la base de datos abierta del Instituto Nacional de Evaluación Educativa[2] que presenta información sobre proyecto de evaluación como Quiero Ser Directivo, Homologación, Ser Profesional, Quiero Ser Asesor Auditor, Ser Maestro Recategorización DMQ, Quiero Ser Maestro Intercultural Bilingüe, Quiero Ser Maestro, Ser Maestro, Ser Bachiller, Ser Estudiante en la Infancia, Ser Estudiante Bachillerato Técnico y Ser Estudiante (Ineval, 2021a; Ineval, 2021b; Ineval, 2021c).

[2] http://evaluaciones.evaluacion.gob.ec/BI/bases-de-datos/

Equiparación utilizando TRI en el contexto de una combinación de los diseños NEAT y NEC: Una combinación de los diseños NEAT y NEC se presenta cuando se tienen diseños de equiparación, donde los grupos son no equivalentes con covariables e ítems ancla. Al respecto, Fuentealba (2020), propone los siguientes pasos para lograr una equiparación en este contexto:

1. Calibración de la forma X e Y con covariables mediante regresión latente (Adams y Wu, 2007; Adams, Wilson y Wu, 1997).
2. Calibración concurrente de las formas X e Y (Wingersky y Lord, 1984).
3. Calibración de la forma X con parámetros de los ítems fijos obtenidos del paso dos (Kim, 2006). Los ítems fijos son aquellos ítems considerados como anclas.
4. Obtención de constantes de transformación lineal, utilizando el método de momentos (ver la sección Métodos de momentos para estimar coeficientes de equiparación) y el enfoque de valores plausibles.
5. Escalamiento de la forma Y a la escala X (ver la sección Linking de parámetros) utilizando los resultados del paso anterior.

El desarrollo de los pasos 1 2, 3 y 4 se explican en los apartados Calibración de la forma X e Y con covariables mediante regresión latente, Equiparación de las puntuaciones de TRI mediante Calibración de parámetros de ítems fijos, Métodos de momentos para estimar coeficientes de equiparación y Linking de parámetros, respectivamente. Estos pasos anteriores se pueden desarrollar con la utilización del paquete TAM del software estadístico R.

Resultados

Después de indicar la metodología pertinente para la realización de una equiparación entre dos formas de prueba utilizando TRI en el contexto de una combinación de los diseños NEAT y NEC, y de describir en el apartado de introducción la parte teórica de los 5 pasos necesarios para llevar a cabo la equiparación, a continuación se presenta el resultado principal de esta investigación que corresponde al código en software estadístico R paso a paso para desarrollar una equiparación en este contexto.

Información necesaria

En primer lugar, se necesita para describir la metodología i) matriz 0/1 de la forma de prueba X, es decir, la matriz de calificación de los sustentantes que desarrollaron la forma de prueba X, ii) matriz 0/1 de la forma de prueba Y, iii) matriz 0/1 compuesta por la combinación de la matriz 0/1

de *X* y la matriz 0/1 de *Y*, el aspecto de esta última matriz se puede ver en la Tabla 1, iv) matriz de factores asociados de los sustentantes de la forma de prueba *X* y v) matriz de factores asociados de los sustentantes de la forma de prueba *Y*. Como la idea es mostrar el desarrollo del análisis de forma general a continuación se expone los datos y el código necesario para simular estas matrices.

```
N_X <- 500 # número de sustentantes en la forma X
N_Y <- 500 # número de sustentantes en la forma Y
I_X <- 10 # número de ítems únicos en la forma X
I_Y <- 10 # número de ítems únicos en la forma Y
I_A <- 10 # número de ítems ancla en las formas X e Y
Y_X_v1_media <- 0 # media de la distribución en la covariable 1 de la forma X
Y_X_v1_desv <- 1 # desviación estándar de la distribución en la covariable 1 de la forma X
Y_X_v2_media <- 0 # media de la distribución en la covariable 2 de la forma X
Y_X_v2_desv <- 1 # desviación estándar de la distribución en la covariable 2 de la forma X
Y_Y_v1_media <- 0 # media de la distribución covariable 1 de la forma Y
Y_Y_v1_desv <- 1 # desviación estándar distribución covariable 1 de la forma Y
Y_Y_v2_media <- 0 # media distribución covariable 2 de la forma Y
Y_Y_v2_desv <- 1 # desviación estándar distribución covariable 2 de la forma Y
theta_X_media <- -2 # media de la distribución theta en la población de la forma X
theta_X_sd <- 1 # desviación estándar de la distribución theta en la población de la forma X
theta_Y_media <- 2 # media de la distribución theta en la población de la forma Y
theta_Y_sd <- 1 # desviación estándar de la distribución theta en la población de la forma Y
n <- 30 # Número de valores plausibles tomados aleatoriamente de la distribución de habilidades de los sustentantes
```

Mediante la información anterior y las siguientes funciones escritas en R se puede simular las matrices necesarias para el proceso de equiparación.

```
Probabilidad <- function(theta, b, a = 1, c = 0){
  c + (1-c)/(1+exp(-a*(theta-b)))
} # Probabilidad de responder correctamente un ítem con parámetros b, 1 y 0 de un sustentante con habilidad theta
calificación <- function(theta, b, a = rep(1, length(b)), c = rep(0, length(b))){
  rbinom(length(b),1,Probabilidad(theta, b, a, c))
```

```
} # Simulación 0/1, es decir, simulación de respuesta incorrecta (0) o correcta (1)
data <- function(Theta, ItemPar){
  N <- length(Theta)
  res <- NULL
  for (i in 1 : N)
    res <- rbind(res, calificacion(Theta[i],ItemPar[1, ], ItemPar[2, ], ItemPar[3, ]))
  return(res)
} # Simulación de la matriz 0/1, con los sustentantes en las filas y los ítems en las columnas
```

Ahora, se puede obtener las 5 matrices necesarias del siguiente modo:

```
# Simulación de la dificultad en los ítems
bX <- rnorm(I_X, -2, 1)
bY <- rnorm(I_Y, 2, 1)
bA <- rnorm(I_A, -2, 1)
# Simulación de theta (habilidad de los sustentantes) utilizando las covariables
theta_X <- rnorm(N_X, mean = theta_X_media, sd = theta_X_sd) + 0.2 * Y_X[,1] + 0.1 * Y_X[,2]
theta_Y <- rnorm(N_Y, mean = theta_Y_media, sd = theta_Y_sd) + 0.2 * Y_Y[,1] + 0.1 * Y_Y[,2]
# Simulación de matrices 0/1 de la forma X, Y combinada
p_X <- rbind(c(bX,bA),1,0)
p_Y <- rbind(c(bA,bY),1,0)
#set.seed(9568)
matriz_X <- data(theta_X, p_X)
matriz_Y <- data(theta_Y, p_Y)
matriz_X_Y <- matrix(NA,nrow = N_X + N_Y, ncol = I_X + I_A + I_Y)
matriz_X_Y[1:N_X, 1:(I_X + I_A)] <- matriz_X
matriz_X_Y[(N_X + 1):(N_X + N_Y), (I_X + 1):(I_X + I_A + I_Y)] <- matriz_Y
matriz_X_Y <- data.frame(matriz_X_Y)
# Nombre de los ítems
item_X <- paste("X", 1:I_X, sep="")
item_A <- paste("A", 1:I_A, sep="")
item_Y <- paste("Y", 1:I_Y, sep="")
colnames(matriz_X_Y) <- c(item_X, item_A, item_Y)
```

matriz_X_Y

 # Simulación de Covariables

Y_X <- cbind(v1=rnorm(N_X, mean = Y_X_v1_media, sd = Y_X_v1_desv), v2=rnorm(N_X, mean = Y_X_v2_media, sd = Y_X_v2_desv))

Y_Y <- cbind(v1=rnorm(N_Y, mean = Y_Y_v1_media, sd = Y_Y_v1_desv), v2=rnorm(N_Y, mean = Y_Y_v2_media, sd = Y_Y_v2_desv))

Ya que se cuenta con los datos, funciones en R y matrices necesarias para el desarrollo de una equiparación entre dos formas de prueba utilizando TRI en el contexto de una combinación de los diseños NEAT y NEC. Se muestra en seguida paso a paso el modo de realizar la equiparación en R.

Paso 1: Calibración de la forma X e Y con covariables mediante regresión latente (Adams y Wu, 2007; Adams, Wilson y Wu, 1997).

Calibración forma X con covariables ----

modI1 <- TAM::tam.mml(resp = matriz_X, Y = Y_X, beta.fixed = FALSE) # Calibración unidimensional con regresores latentes

modI1$xsi # Vector de dificultades de los ítems y sus correspondientes errores estándar

wmodI1 <- TAM::tam.wle(modI1); wmodI1$theta # calcula los parámetros de los sustentantes

modI2 <- tam.pv(modI1, nplausible = n); VP_I1 <- modI2$pv; VP_I1 # Valor plausible, es una representación del rango de habilidades que tienen cada sustentante (estos valores se toman aleatoriamente de la distribución de resultados o habilidades)

FIN Calibración forma X con covariables ----

Calibración forma Y con covariables ----

modf1 <- TAM::tam.mml(resp = matriz_Y, Y = Y_Y, beta.fixed = FALSE)

modf1$xsi

wmodf1 <- TAM::tam.wle(modf1); wmodf1$theta

modf2 <- tam.pv(modf1, nplausible = n); VP_F <- modf2$pv; VP_F

FIN Calibración forma Y con covariables ----

Paso 2: Calibración concurrente de las formas X e Y (Wingersky y Lord, 1984).

modRasch.cc <- tam.mml(resp = matriz_X_Y, Y = NULL, irtmodel = "1PL"); modRasch.cc$xsi # Vector de dificultades de los ítems y sus correspondientes errores estándar

wmodRasch.cc <- TAM::tam.wle(modRasch.cc); wmodRasch.cc$theta # cálculo de los parámetros del sustentante

modRasch.cc2 <- tam.pv(modRasch.cc, nplausible = n); VP_cc <- modRasch.cc2$pv; VP_cc # Valor plausible, es una representación del rango de habilidades que tienen cada sustentante (estos valores se toman aleatoriamente de la distribución de resultados o habilidades)

Paso 3: Calibración de la forma X con parámetros de los ítems fijos obtenidos del paso dos (Kim, 2006). Los ítems fijos son aquellos ítems considerados como anclas.

xsi0 <- modRasch.ccxsixsi[1:(I_X + I_A)]; xsi.fixed <- cbind(1:(length(xsi0)), xsi0) # Calibración de la forma de prueba X con parámetros de los ítems fijos obtenidos del paso dos

mod.xsi.fixed1 <- TAM::tam.mml(resp = matriz_X, irtmodel = "1PL", xsi.fixed = xsi.fixed); mod.xsi.fixed1$xsi # Vector de dificultades de los ítems y sus correspondientes errores estándar

wmod.xsi.fixed1 <- TAM::tam.wle(mod.xsi.fixed1); wmod.xsi.fixed1$theta # cálculo de los parámetros del sustentante

mod.xsi.fixed2 <- tam.pv(mod.xsi.fixed1, nplausible = n); VP_I2 <- mod.xsi.fixed2$pv; VP_I2 # Valor plausible, es una representación del rango de habilidades que tienen cada sustentante (estos valores se toman aleatoriamente de la distribución de resultados o habilidades)

Paso 4: Obtención de constantes de transformación lineal, utilizando el método de momentos y el enfoque de valores plausibles.

VP_I1 <- as.matrix(VP_I1[,-1])

VP_I2 <- as.matrix(VP_I2[,-1])

B <- sd(VP_I1) / sd(VP_I2); B

*A <- mean(VP_I1) - B * mean(VP_I2); A*

Paso 5: Escalamiento de la forma Y a la escala X utilizando los resultados del paso anterior.

VP_F <- apply(VP_F[,-1], 1, median)

*VP_F_equiparado <- A + B * VP_F*

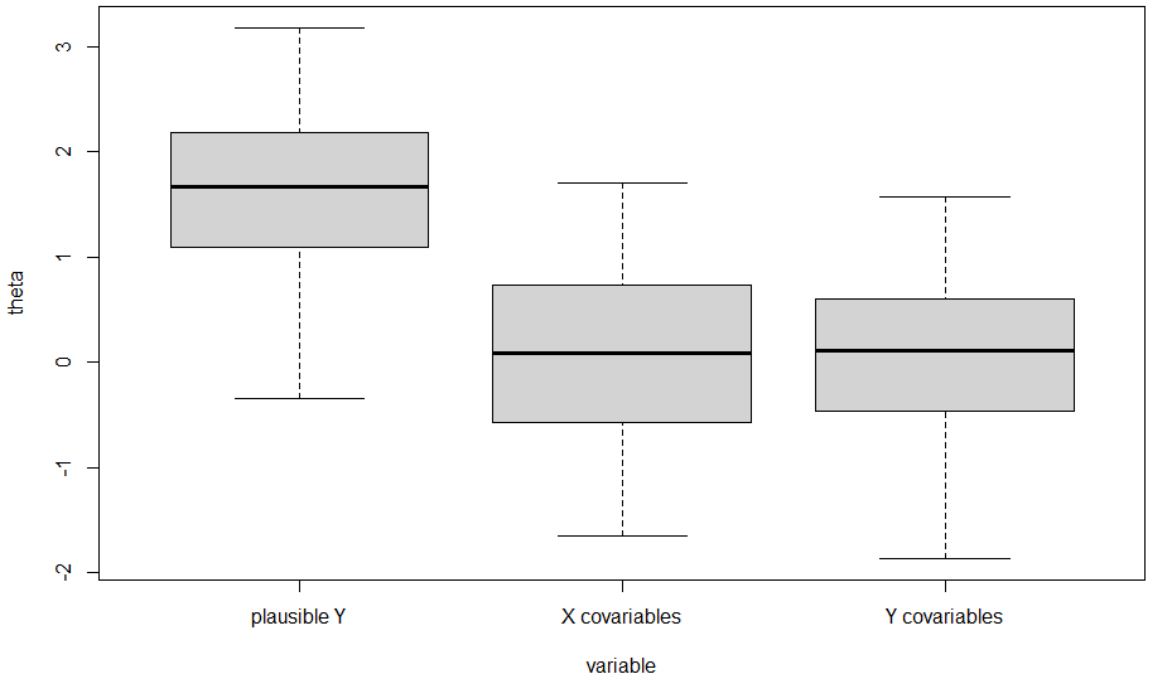
Luego de ejecutar estos cinco pasos en el software estadístico R, se consigue desarrollar una equiparación entre dos formas de prueba utilizando TRI en el contexto de una combinación de los diseños NEAT y NEC. Para finalizar la sección de resultados se cree que es conveniente graficar

la distribución de las habilidades (theta) predichas por esta metodología[3], mediante un diagrama de caja (ver Figura 1 y Tabla 2, el código necesario para realizar la Figura y la tabla se presentan más adelante). Este diagrama evidencia la importancia de realizar una equiparación al comparar las habilidades de los sustentantes, debido a que, si no se hubiera realizado la equiparación, parecería que las habilidades de los sustentantes de las formas de prueba *X* (*X* covariables) e *Y* (*Y* covariables) son muy similares, sin embargo, esto no es verdad ya que al desarrollar la equiparación se muestra que en realidad las habilidades de los sustentantes de la forma de prueba *Y* (plausible *Y*) es más alta que la de los sustentantes de la forma de prueba *X*.

```
thetas_X_cov <- apply(VP_I1[,-1], 1, median)
thetas_Y_cov <- VP_F
thetas_plausible <- VP_F_equiparado
resumen_f <- data.frame(variable = c(rep("X covariables", length(thetas_X_cov)),
                                     rep("plausible Y", length(thetas_plausible)),
                                     rep("Y covariables", length(thetas_Y_cov)) ),
                        theta = c(thetas_X_cov, thetas_plausible, thetas_Y_cov) )
boxplot(theta ~ variable, data = resumen_f)
Estadisticos <- data.frame(Estadístico = c("Minimo", "1er cuartil", "Mediana", "Media", "3er cuartil", "Máximo", "Desv.estar"),
                           thetas_X_cov = c(summary(thetas_X_cov), sd(thetas_X_cov)),
                           thetas_Y_cov = c(summary(thetas_Y_cov), sd(thetas_Y_cov)),
                           thetas_plausible_Y = c(summary(thetas_plausible), sd(thetas_plausible)) )
```

[3] Aquí se muestra una ejecución del código, sin embargo, el gráfico puede cambiar debido a que se ha desarrollado una simulación sin considerar el valor de semillas.

Figura 1. Comparación entre las distribuciones de habilidad predichas en las formas de prueba *X* e *Y*



Elaborado: Autores

Tabla 2. Estadísticas descriptivas de las habilidades predichas de las formas de prueba *X* e *Y*

Estadístico	X covariable	Y covariable	plausible Y
Mínimo	-2.0818	-2.0351	-0.3473
1er cuartil	-0.6040	-0.6111	1.0635
Mediana	0.0100	0.0840	1.7522
Media	-0.0136	0.0045	1.6734
3er cuartil	0.6318	0.5971	2.2606
Máximo	1.9989	1.7419	3.3949
Desviación Estándar	0.8908	0.8340	0.8263

Fuente: Autores, 2022

Conclusiones

En el contexto de las evaluaciones estandarizadas a gran escala en el Ecuador, a menudo es necesario comparar los resultados de los sustentantes que rindieron dos formas de prueba diferentes, esto se puede realizar mediante una equiparación entre dos formas de prueba utilizando TRI en el contexto de una combinación de los diseños NEAT y NEC. Para el desarrollo con éxito de la equiparación se puede ocupar la siguiente metodología de 5 pasos implementada en la sección

de resultados en el software estadístico R: i) Calibración de la forma X e Y con covariables mediante regresión latente; ii) Calibración concurrente de las formas X e Y; iii) Calibración de la forma de prueba X con parámetros de los ítems fijos obtenidos del paso dos, donde los ítems fijos son aquellos ítems considerados como anclas; iv) Obtención de constantes de transformación lineal, utilizando el método de momentos y el enfoque de valores plausibles y v) Escalamiento de la forma de prueba Y a la escala X utilizando los resultados del paso anterior.

Finalmente, es importante indicar dos cuestiones: i) cuando se pretenda compara las habilidades de los sustentantes que rindieron dos formas de prueba diferentes, siempre se debe realizar una equiparación para que sean comparables las diferencias en la media y la desviación estándar de las distribuciones de habilidad estimadas de las diferentes formas de prueba, tomado a una de ellas como escala base, en el ejemplo presentado en este trabajo la escala base fue la forma de prueba X y ii) si se desea comparar los resultados de más de dos formas diferentes de prueba es necesario realizar un proceso de encadenamiento, es decir, una transformación secuencia entre formas de prueba contiguos.

Referencias

1. Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model. A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.): Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 55-76). New York: Springer. Doi: 10.1007/9780387498393_4
2. Adams, R. J., Wilson, M. R., and Wu, M. L. (1997). Multilevel item response modelling: An approach to errors in variables regression. Journal of Educational and Behavioral Statistics, 22, 47-76.
3. Burke, K. (1999). The mindful school: How to assess authentic learning (3rd Ed.). Arlington Heights, IL: Skylight Publishing.
4. DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. Applied Measurement in Education, 15(1), 15–31.
5. Fuentealba, F. (2020). Sesión 3 – ERCE y Rosseta Stone, ERCE – Metodología y Aplicación. Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.

6. González, J., & Wiberg, M. (2017). Applying test equating methods. Cham: Springer International Publishing.
7. Ineval. (2021a). Ficha técnica, Ser Estudiante, Cuarto de Educación General Básica. http://evaluaciones.evaluacion.gob.ec/BI/ser-estudiante/
8. Ineval. (2021b). Acción 21-3: Capacidad en la resolución de problemas en los sistemas informatizados, de las personas que participaron en la tercera ronda del Programa para la Evaluación Internacional de las Competencias de los Adultos (PIAAC). http://evaluaciones.evaluacion.gob.ec/BI/ser-estudiante/
9. Ineval. (2021c). Acción Boletines de Investigación y Evaluación: Involucramiento parental y rendimiento académico en escolares de 7mo de Educación General Básica.
10. Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. Journal of Educational Measurement, 43(4), 355–381.
11. Kolen MJ, Brennan RL. Test Equating, Scaling, and Linking: Methods and Practices. New York, NY: Springer; 2014.
12. Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley; 1968.
13. Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. Psychometrika, 54, 661-679.
14. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago, IL: University of Chicago Press; 1960.
15. Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8(3), 347–364.