SABER AL DETALLE

EDICIÓN
03
Bogotá D.C.
Enero de 2019
ISSN: 2590-4663
Publicación trimestral

Instituto Colombiano para la Evaluación de la Educación, ICFES
Oficinas: Calle 26 NO. 69-76 Torre 2, piso 15
Edificio Elemento, Bogotá • Colombia

Directora General: **María Figueroa Cahnspeyer**
Directora de Evaluación: **Natalia González Gómez**
Subdirectora de Análisis y Divulgación: **Ana María Restrepo Sáenz**

Subdirección de Estadísticas: **Jorge Mario Carrasco Ortiz**
Subdirección de Diseño de Instrumentos: **Javier Toro Baquero**
Coordinación General: **Dirección de Evaluación**

¿QUÉ GARANTIZA LA COMPARABILIDAD DE LOS RESULTADOS EN LAS PRUEBAS SABER REALIZADAS POR EL ICFES?

¿QUÉ GARANTIZA LA COMPARABILIDAD DE LOS RESULTADOS EN LAS PRUEBAS SABER REALIZADAS POR EL ICFES?

Para garantizar la comparabilidad de las mediciones de las competencias de los evaluados de una prueba que es aplicada a diferentes poblaciones y en distintos períodos, se realiza un procedimiento llamado equiparación. Este permite hacer comparables las mediciones entre diferentes aplicaciones de una misma prueba y, por tanto, hace posible evaluar cómo evoluciona la calidad de la educación, de acuerdo con la periodicidad que maneja el ICFES para la aplicación de cada una de las pruebas (en el caso de Saber 3°, 5° y 9° es anual, en Saber 11 es semestral, en Saber TyT es semestral y en Saber PRO es anual). Como se verá más adelante, en la literatura existen diferentes métodos de equiparación y para la escogencia del mismo es relevante considerar el diseño que se usa para el armado de la prueba, ya que esto depende, en buena parte, del método de equiparación a utilizar. En el caso de las pruebas Saber, la construcción de la prueba se hace bajo un diseño en Bloques Incompletos Balanceados (BIBs)[1] , lo que garantiza que la totalidad de los ítems utilizados en una medición han sido previamente piloteados y, por lo tanto, constituyen ítems comunes entre pruebas, también llamados ítems de anclaje.

1. ¿Cómo se garantiza la comparabilidad dentro de una misma aplicación?

Como se menciona en el documento sobre el armado[2] de las pruebas, por construcción, la aplicación de cada prueba está compuesta por formas (es decir, agrupación de bloques), las cuales tienen ítems comunes. La manera como se estiman las características (parámetros) de los ítems determina la necesidad, o no, de llevar a cabo un proceso de equiparación. Si las estimaciones se realizan con la juntura de todas las formas, los parámetros van a estar en la misma métrica y por tanto la comparabilidad es directa. De otra forma, si la estimación de dichas aplicaciones se realiza separada por formas, las estimaciones de los ítems van a estar en escalas diferentes, lo que hace imprescindible el uso de un proceso de equiparación para que las habilidades se encuentren en misma escala. En el caso de las pruebas Saber, para garantizar la comparabilidad de la prueba dentro de una misma aplicación, se realiza un proceso de estimación conjunta que permite establecer una misma métrica para las estimaciones de los parámetros de los ítems y dado esto, establecer una métrica común para la estimación de las habilidades de los evaluados.

Para que este proceso ocurra de manera directa, es importante garantizar que las poblaciones que presentan las formas sean equivalentes y que las formas sean homogéneas. Por un lado, las formas de una prueba se construyen bajo unas mismas especificaciones estadísticas y de contenidos; esto se conoce como "formas de prueba nominalmente paralelas" (Lord and Novick, 1968), de tal manera que se garantiza que todas las formas apuntan hacia el mismo constructo que se busca medir y lo hacen bajo características estadísticas homogéneas. Por su parte, para garantizar que las poblaciones sean equivalentes, se hace una asignación aleatoria de las formas entre los estudiantes que presentan la prueba. Esta asignación aleatoria, dentro de un mismo período, garantiza que los estudiantes que presentan cada una de las formas sean similares en términos de sus características y, por lo tanto, no hay diferencias sistemáticas entre ellos que generen diferencias en sus habilidades.

1 Para profundizar el diseño del armado y sus supuestos refiérase a la Edición 2 del Boletín Saber al Detalle "¿Qué diseño del armado se emplea en el Icfes para medir las pruebas Saber?"

2 *Ibid.*

2. ¿Cómo se garantiza la comparabilidad entre aplicaciones?

Al aplicar una prueba en distintos períodos, en la práctica no puede garantizarse que los grupos de evaluados compartan las mismas características, por lo cual en cada aplicación los grupos no son equivalentes y los parámetros de los ítems pueden presentar diferencias. Precisamente, para ajustar estas diferencias mencionadas se realiza el proceso de equiparación que permite comparabilidad de puntajes, a pesar de no haber presentado las mismas formas de la prueba (Kolen and Brennan, 2014).

Recordemos que para las pruebas Saber se estiman las habilidades de los evaluados empleando un modelo de calificación de teoría respuesta al ítem (TRI) de 3 parámetros (3PL)[3]. Teniendo en cuenta que el grupo de evaluados varía entre aplicaciones, la estimación de los parámetros de los ítems puede presentar diferencias. Por ello, se hace necesario emplear un método estadístico que permita generar comparabilidad entre aplicaciones. El proceso de equiparación es ese método estadístico que permite garantizar dicha comparabilidad y la lógica que se usa es la de definir una métrica común entre aplicaciones, que permita hacer comparaciones directas.

3. ¿Cuáles son los métodos de equiparación?

Existen básicamente tres métodos de equiparación basados en TRI que permiten tener una escala común de los puntajes de las aplicaciones. Para ello, se contemplan dos tipos de aproximaciones de diseño de recolección de datos: grupos comunes de evaluados y/o grupos no equivalentes de evaluados con test de anclaje (NEAT por sus siglas en inglés). En el primer caso, hay personas en común entre las diferentes aplicaciones; en el segundo caso se contemplan grupos distintos de evaluados y se asume que el grupo de ítems de anclaje puede cuantificar la diferencia de habilidades entre ellos. En el caso de las pruebas Saber, se emplea la segunda aproximación. A continuación, se describen los métodos de equiparación:

1. Transformaciones lineales de calibraciones separadas

Bajo este método de equiparación, los parámetros de los ítems de varias aplicaciones de una prueba se calibran por separado para cada período de aplicación. Luego, bajo un método que depende de cada técnica, se calculan los coeficientes de transformación para realizar el proceso de equiparación, los cuales se derivan de los ítems comunes entre aplicaciones. A partir de dichas constantes de equiparación, se realiza una transformación lineal los parámetros de todos los ítems. Dichas transformaciones lineales pueden ser estimadas a través de los métodos Mean/Sigma, Mean/Mean, Haebara o Stocking y Lord[4]. Aunque para cada método varía la forma en que se calculan las constantes de transformación, cada método busca minimizar diferencias existentes entre las estimaciones de los parámetros o de los promedios y desviaciones de las habilidades entre los grupos. Con respecto a las dos transformaciones no hay un consenso sobre cual método es mejor, por lo cual se recomienda considerarlos y comparar las escalas de la aplicación (Kolen and Brennan, 2014).

2. Calibraciones con parámetros fijos de ítems comunes (FCIP, por sus siglas en inglés)

Bajo este método, los parámetros de los ítems que son comunes con aplicaciones anteriores se fijan. El objetivo de este enfoque consiste en calibrar los ítems nuevos, teniendo fijos los parámetros de los ítems de anclaje, para que se encuentren en la misma escala y con esto definir una misma métrica que garantice la comparabilidad.

Este tipo de metodologías resulta muy relevante bajo un escenario en el cual el diseño del armado de una prueba genera formas comparables y donde existen ítems comunes que se emplean entre aplicaciones y, por tanto, es conocido el comportamiento psicométrico de los mismos. En tal caso, el reto consiste en contemplar las diferencias sutiles en la distribución de habilidades a raíz de la aplicación continua de una misma prueba que en agregado pueda afectar la estimación de habilidades. Se ha encontrado que para equiparación en múltiples

3 Para profundizar el proceso calibración y sus supuestos refiérase a la Edición 1 del Boletín Saber al Detalle "¿Cómo se generan los puntajes en las pruebas Saber del Icfes?"

4 Dichos métodos se pueden encontrar en Arai and Mayekawa (2011)

aplicaciones de una prueba, la metodología FCIP no se afecta de la misma manera que los métodos de transformación (Keller and Keller, 2011). Dado el armado de las pruebas Saber, el cual garantiza que todos los ítems son comunes con aplicaciones anteriores, este método es ideal para garantizar la comparabilidad.

3. Calibración conjunta

Bajo esta metodología, todos los parámetros de los ítems de las aplicaciones de una prueba son estimados a través de una sola calibración. Es decir, que se incluyen todos los ítems de las aplicaciones y todos los individuos y se realiza un solo proceso de estimación de los parámetros de los ítems, por lo cual todos los ítems están bajo la misma escala. A la hora de emplear la calibración conjunta es necesario definir a qué aplicación hace parte cada evaluado, para tener en cuenta diferencias en la habilidad entre grupos (DeMars, 2002). Se ha encontrado que una desventaja de la calibración conjunta tiene que ver con el análisis por parámetros estimados de los ítems comunes, ya que bajo la calibración conjunta se estiman los parámetros de todos los ítems comunes a la vez y no de manera individual (Kolen and Brennan, 2014). Otra desventaja, es que este método no es práctico en los casos de aplicaciones repetitivas, dado que en cada nueva aplicación se debería repetir el proceso de estimación y por tanto cambiaría la estimación de las habilidades de los evaluados. Este es el caso de las pruebas Saber, las cuales son presentadas semestral o anualmente y por lo tanto este método no es viable para hacer la equiparación.

4. ¿Cómo funciona la equiparación en la práctica?

Con la aplicación de las pruebas Saber se tiene como objetivo, entre otros, (1) estimar las habilidades de los evaluados, (2) hacer comparaciones entre periodos para analizar cómo evoluciona la calidad de la educación entre grados educativos y (3) pilotear ítems nuevos para obtener la estimación de sus parámetros y ser utilizados en aplicaciones posteriores. Así, en el armado de cada prueba se incluyen ítems nuevos (pilotos) e ítems que han sido medidos en otras aplicaciones. En la figura 1 se ilustra un ejemplo del armado de una prueba Saber 11 para el segundo periodo de 2018: está compuesta por ítems de medición aplicados en los periodos 2016-1, 2016-2, 2017-1, 2017-2 e ítems piloto. Los ítems comunes entre dichos periodos y la aplicación 2018-2 sirven de ancla para conseguir una escala común y centrar las habilidades en la aplicación de la línea base: como han sido previamente calibrados en periodos anteriores se conoce su comportamiento.

En particular, se puede apreciar que los bloques de ítems "P1 2016" y "P2 2016" que hacen parte del armado en la aplicación del periodo 2018-2 fueron piloteados en los dos periodos del 2016 y, una vez conocido el comportamiento de los parámetros, a partir del primer semestre de 2017 se emplearon como ítems de medición. De esta manera se ilustra cómo a través del pilotaje de nuevos ítems en las pruebas se genera el proceso de calibración para ítems que van a ser empleados en posteriores aplicaciones. Un supuesto que se tiene en este proceso es que no existe comportamiento diferencial en los ítems (conocido como DIF, por sus siglas en inglés). Este tema será tratado en un próximo documento.

Figura 1. Bloques de ítems de anclaje e ítems pilotos para una prueba Saber 11 histórico

	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2
Items de mediación	P1 2015 P2 2015 P3 2015 P4 2015	P1 2014 P2 2014 P3 2015 P4 2015	P1 2016 P2 2016 P3 2016 P4 2016	P1 2015 P2 2015 P3 2016 P4 2016	P1 2017 P2 2017 P3 2017 P4 2017	P1 2016 P2 2016 P3 2017 P4 2017
Items pilotos	P2 2016 P3 2016 P4 2016	P1 2016	P2 2017 P3 2017 P4 2017	P1 2017	P2 2018 P3 2018 P4 2018	P1 2018

Fuente: ICFES, 2018

Se ha encontrado que al aplicar cualquiera de las metodologías anteriores, bajo condiciones óptimas, los tres métodos arrojarían resultados similares (Kang and Petersen, 2009). Dado el sustento teórico de los tres métodos de calibración, se espera que las diferencias que podría haber entre la aplicación de ellos sean pequeñas (Jodoin et al., 2003). De igual manera, se recomienda en la práctica implementar cada uno de los métodos y evaluar los efectos de emplear una u otra sobre la comparabilidad entre aplicaciones.

Bibliografía

Arai, S. and Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviourmetrika,* 38(1):1-16.

DeMars, C. (2002). Incomplete data and item parameter estimates under jmle and mml. *Applied Measurement in Education,* 15:15-31.

ICFES (2017). ¿Cómo se generan los puntajes en las pruebas saber del icfes?. *Sin publicar.*

ICFES (2018). ¿Qué diseño se emplea de armado se emplea en el icfes para medir las pruebas saber? *Sin publicar.*

Jodoin, M., Keller, L., and Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71(3):229-250.*

Kang, T. and Petersen, N. (2009). *Linking Item Parameters to a Base Scale,* volume 2.ACT Research Report Series.

Keller, L. and Keller, R. (2011). The long-term sustainability of diferent item response theory scaling methods. *Educational and Psychological Measurement,* 71(2):362-379.

Kolen, M. and Brennan, R. (2014). *Test Equating, Scaling, and Linking: Methods and Practices. Springer.*

Lord, F. and Novick, M. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Weller, S. and Romney, A. (1988). *Systematic Data Collection Vol 10.* Sage Publications.

SABER AL DETALLE

¿QUÉ GARANTIZA LA COMPARABILIDAD DE LOS RESULTADOS EN LAS PRUEBAS SABER REALIZADAS POR EL ICFES?

icfes mejor saber

La educación es de todos
Mineducación

@icfescol | ICFES | icfescol | YouTube: ICFES

Instituto Colombiano para la Evaluación de la Educación, ICFES
Oficinas: Calle 26 NO. 69-76 Torre 2, piso 15
Edificio Elemento, Bogotá • Colombia