Edición 25
Bogotá D.C.
Diciembre de 2017
ISSN: 500-445X

Boletín SABER en breve

Instituto Colombiano para la Evaluación de la Educación, ICFES.
Oficinas: Calle 26 No.69-76, Torre 2, Piso 15, Edificio Elemento,
Bogotá D.C., Colombia.

EL ARMADO DE LAS PRUEBAS SABER Y LA COMPARABILIDAD EN EL TIEMPO

El armado de las pruebas Saber y la comparabilidad en el tiempo

En este Saber en Breve describimos el procedimiento con el que ensamblamos los cuadernillos y armamos las pruebas Saber. Al momento de la aplicación del examen, los estudiantes evaluados presentan un conjunto de preguntas de la prueba. Estos conjuntos tienen una dificultad promedio y un rango de dificultades comparables con otros conjuntos de la prueba en la misma aplicación y entre aplicaciones. De esta forma, brindamos resultados comparables en el tiempo.

El mundo ideal

En el Icfes garantizamos la comparabilidad de los resultados de las pruebas en el tiempo. De esta forma, el resultado de un estudiante que presentara la prueba varias veces debería ser aproximadamente el mismo independiente del periodo de aplicación (a menos que el estudiante tenga un cambio). La forma más fácil de garantizar este objetivo sería aplicando la misma prueba. Si las preguntas del examen son las mismas, por definición la dificultad y el rango de dificultad de los exámenes es el mismo.

Los resultados comparables permiten cuantificar dónde se dan los cambios, por ejemplo, es posible saber cuáles colegios del país tienen mayor desempeño o mejoraron más respecto del año pasado y en qué componentes de la prueba tienen mayor dominio. Y lo más importante: se puede cuantificar la magnitud de las diferencias o los cambios. Es decir, no solamente sabemos cuáles colegios y en qué componentes de la prueba, sino que también sabemos en cuánto cambiaron. Una región del país pudo mejorar, por ejemplo, tanto en matemáticas como en lenguaje, pero pudo mejorar más en un área. Saber la magnitud de los cambios permite un diagnóstico de los aprendizajes que es útil para aprender de la evaluación.

Exámenes diferentes

Aunque por comparabilidad sería deseable indagar las mismas preguntas, hay dos razones principales por las cuales no aplicamos exámenes con los mismos ítems. La primera es que, a excepción de los que tomen el examen en la primera aplicación (estudiantes), las personas podrían aprenderse de memoria algunas preguntas y se cambiaría el objetivo de la evaluación por competencias por una evaluación de qué tan buena memoria tienen los evaluados (además, podría aumentar la probabilidad de copia). La segunda razón es que el examen de cada estudiante sería considerablemente largo porque cada uno tendría que responder todas las preguntas que conforman la totalidad del examen y no un conjunto como lo mencionamos al principio.

Cabe aclarar que, aunque los exámenes, tanto dentro como entre aplicaciones, son diferentes, una proporción de la prueba se compone de ítems de anclaje. Estos ítems son comunes entre diferentes formas y aplicaciones y sirven para facilitar el objetivo de hacer la medición comparable.

Los bloques y las formas de los exámenes

Las preguntas que conforman una prueba o ítems de los exámenes los dividimos en bloques que tienen dificultades promedio iguales y variación de la dificultad de los ítems iguales (es decir, el promedio y la desviación estándar de la dificultad de cada bloque es la misma). Además, cada bloque tiene una proporción fija de las competencias y componentes evaluados en cada materia. Grupos de bloques conforman formas de medición (cuadernillos) que presentan los estudiantes evaluados.

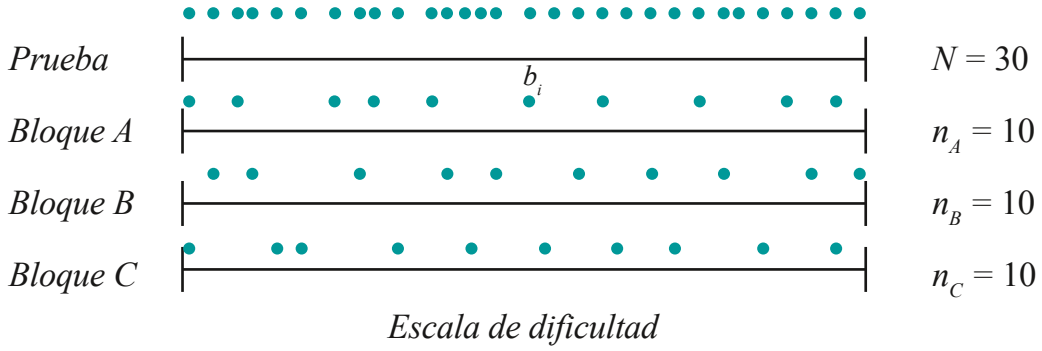
Los Bloques Incompletos Balanceados (BIBs)

Los Bloques Incompletos Balanceados son un esquema de rotación de cuadernillos, de anclaje y de pilotaje que permite blindar este proceso desde el punto de vista técnico y garantizar la comparabilidad.

La Figura 1 es una representación de la posición relativa de 30 ítems que componen una prueba a lo largo del rango de dificultad adecuado para una población. La división en bloques responde al tiempo disponible para la aplicación de la prueba. En este ejemplo, un estudiante responde únicamente un bloque, a partir del cual se infiere su desempeño global. Por lo tanto, aunque existen 30 preguntas, cada estudiante responde solamente 10 y, aunque no todos responden las mismas, es posible comparar sus calificaciones.

La Tabla 1 muestra, como ejemplo, las formas y bloques de la prueba Saber 3°, 5° y 9°. En total, hay 10 bloques, que van del A al J. Cada forma se compone de tres bloques. Así que hay 30 formas de medición. Este procedimiento cumple

Figura 1. Ejemplo de Bloques Incompletos Balanceados



que cada bloque esté el mismo número de veces en total, que una pareja de bloques esté el mismo número de veces dentro de una forma y que cada bloque esté el mismo número de veces en cada posición posible (es decir, de primero, segundo o tercero en el cuadernillo).

Con un diseño de Bloques Incompletos Balanceados (BIBs) no es necesario obtener las respuestas del estudiante en todos los ítems que conforman la prueba. Al tener una gran cantidad de respuestas individuales sobre grupos de ítems, se puede predecir el desempeño individual que habría obtenido un estudiante si hubiera respondido la totalidad de la prueba. Es decir, se maximiza la información que proporcionan las respuestas individuales sobre varios bloques de ítems.

La motivación para usar BIBs responde a las restricciones de tiempo de la aplicación, a la necesidad de estimar una puntuación global individual del desempeño para cada una de las pruebas, estandarizar el diseño para ensamblar distintas pruebas a lo largo de tiempo y maximizar la información que se obtiene de la aplicación de los exámenes. A lo anterior se añade la necesidad de proporcionar una puntuación individual longitudinal con fines de seguimiento y la definición de una escala transversal de desempeño a lo largo de varios ciclos educativos (por ejemplo, para el objetivo de hacer comparable los puntajes de cada grado en la prueba Saber 3°, 5° y 9°).

Las preguntas de evaluación y las preguntas de pilotaje

Para poder decir que los bloques tienen dificultades comparables, las preguntas a incluir en cada bloque deben tener una estimación de su dificultad. Esta información se tiene porque todas las preguntas que usamos para calificación han sido primero piloteadas. Esto quiere decir que ya hay estudiantes que las han respondido. En cada aplicación hay preguntas de calificación y preguntas de pilotaje. Por lo tanto, no todas las preguntas que responden los estudiantes hacen parte de su calificación y las que sí hacen parte ya han sido piloteadas en aplicaciones

Tabla 1. Formas y bloques de la prueba Saber 3°, 5° y 9°

Forma	*Bloque 1*	*Bloque 2*	*Bloque 3*	Forma	*Bloque 1*	*Bloque 2*	*Bloque 3*
F_1	A	D	B	F_{16}	J	A	H
F_2	E	H	D	F_{17}	F	J	C
F_3	I	C	B	F_{18}	E	J	G
F_4	B	G	E	F_{19}	G	H	D
F_5	F	B	A	F_{20}	B	D	G
F_6	B	G	F	F_{21}	G	E	F
F_7	C	D	A	F_{22}	E	I	C
F_8	H	B	I	F_{23}	H	F	C
F_9	C	E	D	F_{24}	C	G	H
F_{10}	A	E	I	F_{25}	H	A	J
F_{11}	D	I	F	F_{26}	I	G	A
F_{12}	J	C	B	F_{27}	D	F	J
F_{13}	F	A	E	F_{28}	J	B	E
F_{14}	D	I	J	F_{29}	I	F	H
F_{15}	A	C	G	F_{30}	G	J	I

anteriores. Las preguntas piloto, además de ser usadas para estimar la dificultad de los bloques de las siguientes aplicaciones, sirven para hallar errores o comportamientos atípicos en los ítems y sus respuestas. Por ejemplo, podemos saber si los distractores (las opciones de respuesta incorrectas) están cumpliendo su objetivo (de distraer, pero no de confundir) o podemos saber si una pregunta estuvo demasiado fácil o difícil (si todos los estudiantes evaluados responden correcta o incorrectamente una pregunta, esta no presentaría variación y no serviría para entender las diferencias entre estudiantes).

La Teoría de Respuesta al Ítem (TRI)

El procedimiento estadístico con el que se califican las pruebas Saber es el modelo logístico de tres parámetros (3PL). Este permite que la evaluación no dependa del conjunto de preguntas que responden los estudiantes (ni del porcentaje de aciertos), sino de la estimación de estadísticos para cada ítem, que permiten armar evaluaciones comparables que se componen de ítems diferentes. Además, los estadísticos estimados no se restringen a la dificultad de los ítems: también permiten conocer la discriminación y la probabilidad de respuestas al azar. Una pregunta con una alta discriminación es una pregunta que separa muy bien los estudiantes según su respuesta: es probable que los que la responden de forma correcta van a tener una calificación total mucho más alta que los que la responden de forma incorrecta. El parámetro de azar muestra que es poco probable que un estudiante responda correctamente una pregunta de alta dificultad si ha respondido incorrectamente preguntas más fáciles de la prueba. La TRI, además de ser utilizada para la calificación, permite tener un banco de ítems calibrados: sirve para tener la información para armar los bloques y formas y para incluir o excluir preguntas piloto en la evaluación (además, permite crear un examen adaptativo por computador).

Después de entender la forma por la cual garantizamos la comparabilidad de las evaluaciones, podemos entender que el hecho de que la evaluación nos permita conocer la procedencia y la magnitud de los cambios no implica que sabemos por qué se generan. Por ejemplo, la mejora de un colegio pudo ser causa de que sus profesores son otros o de que sus estudiantes son otros (y de muchas otras razones que no necesariamente tienen que estar relacionadas con lo que pasa dentro del salón de clase). Pero, desde el Icfes no podemos saber estas razones. Solamente medimos los cambios en un estudio observacional (no experimental). Estas preguntas causales podrían ser respondidas por investigadores con ayuda de los resultados de las pruebas (que ponemos a disposición del público).