SABER AL DETALLE

EDICIÓN
01
Bogotá D.C.
Julio de 2018
ISSN: En trámite
Publicación trimestral

icfes mejor saber

GOBIERNO DE COLOMBIA

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: **Ximena Dueñas Herrera**
Directora de Evaluación: **Natalia González Gómez**
Subdirectora de Análisis y Divulgación: **Silvana Godoy Mateus**

Subdirección de Estadísticas: Edwin Cuellar
Subdirección de Diseño de Instrumentos: Javier Toro
Coordinación General: **Dirección de Evaluación**

¿CÓMO SE GENERAN LOS PUNTAJES EN LAS PRUEBAS SABER DEL ICFES?

¿CÓMO SE GENERAN LOS PUNTAJES EN LAS PRUEBAS SABER DEL ICFES?

Para la calificación de las pruebas Saber 359, Saber 11, Saber TyT y Saber PRO se emplea la Teoría de Respuesta al Ítem (TRI), que permite tomar en cuenta las características psicométricas de las preguntas (también llamadas ítems) y establecer una relación probabilística con la habilidad o trazo latente que se desea medir. El objetivo de la calificación es poder tener una escala definida para medir la habilidad de los evaluados, al ser esta una variable que no se puede medir directamente y que además no tiene una escala únicamente definida. Así, se presentan a continuación los aspectos técnicos que permiten superar estos retos a la hora de generar los puntajes en las pruebas Saber.

Un ejemplo para entender la habilidad o trazo latente sería tratar de medir la estatura de una persona sin un metro. Supongamos que no se cuenta con un instrumento que permita obtener el número correspondiente a la altura, entonces se podría emplear un instrumento con ítems que de indicios de qué tan alta es una persona. Por ejemplo, se podría preguntar: ¿Alcanza a poner una maleta en el equipaje de un avión? ¿Logra sostenerse de los tubos horizontales en un bus? ¿Suele tocar el marco de una pared al estirarse? De esta manera, es posible empezar a estimar la altura de las personas a las que se encueste, con base en sus respuestas a los ítems, sin tener medidas directamente. Esta es la lógica de la TRI.

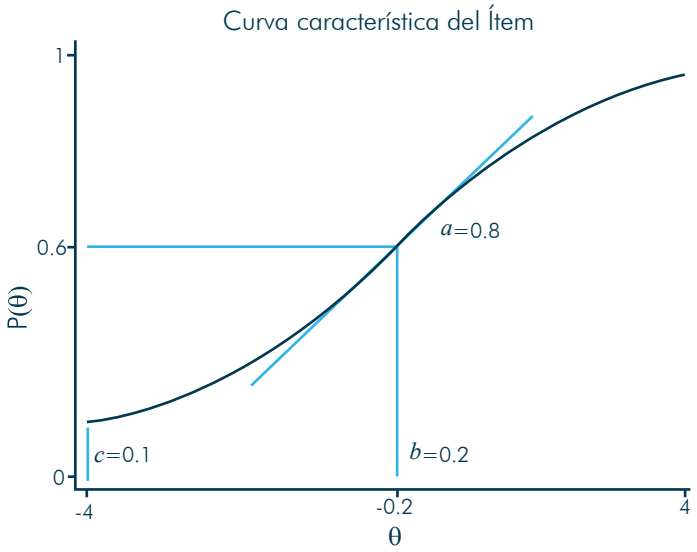
1. ¿Qué significa esto?

La TRI es una teoría de los test para medir variables latentes, es decir que no se pueden medir directamente, a partir de una estimación paramétrica de las respuestas dadas por los evaluados a un conjunto de preguntas o ítems. En el caso de las pruebas Saber, lo que se busca es estimar la habilidad de los evaluados en algunas competencias como razonamiento cuantitativo o lectura crítica, a partir de un conjunto de ítems que miden distintos niveles de habilidad. La TRI permite estimar la habilidad de los individuos (θ) con base en las respuestas dadas a los ítems y las características de los ítems (a los cuales llamamos parámetros). Dichos parámetros pueden ser estimados durante el proceso de estimación (a este proceso lo llamamos calibración) o pueden haber sido calibrados en aplicaciones anteriores de la prueba (a esto lo llamamos anclaje). De esta manera es posible establecer una relación entre habilidad estimada de los evaluados y la probabilidad de acertar el ítem, teniendo en cuenta los parámetros de los ítems.

Actualmente, para las pruebas Saber se emplea un modelo logístico con tres parámetros de los ítems, de allí que se conozca como modelo logístico de tres parámetros (3PL por sus siglas en inglés). Veamos cómo funciona esto en la figura 1. Nos referimos a curva característica de un ítem a la función que representa la probabilidad de acertar el ítem, en términos de la habilidad y de los parámetros del ítem. Esto implica que, dados unos valores de los parámetros del ítem, para cada nivel de habilidad (θ) hay una probabilidad asociada de acierto para ese ítem.

En el eje x se encuentran los valores de la escala de habilidad de los evaluados, es decir de θ, y en el eje y los valores de la probabilidad de acierto del ítem $P(\theta)$. La escala que se emplea para la habilidad (θ) está en el rango de $-\infty$ a ∞, aunque con una probabilidad mayor a 99% se encuentra entre -4 a 4 y, suele normalizarse, de forma que en promedio la habilidad está centrada en cero y se dispersa 1 unidad alrededor de la media. Para el ítem que aparece en la figura 1, los evaluados con una habilidad de $\theta=0.2$ tienen una probabilidad de acierto de 0.6. En la medida en que la habilidad del evaluado disminuye, la probabilidad de acierto es menor y, de manera similar, cuando aumenta la habilidad, aumenta la probabilidad de acertar el ítem. Más adelante explicaremos a qué corresponden los parámetros ilustrados en la figura , los cuales corresponden a: dificultad del ítem (b), discriminación (a) y acierto casual (c).

Figura 1. Curva característica de un ítem en modelo 3PL



Fuente: ICFES, 2017

2. ¿En qué consiste?

Al analizar la relación entre la habilidad y la probabilidad de acierto, cabe señalar que estas no tienen un comportamiento lineal, lo que quiere decir que un aumento en una unidad en habilidad no implica un aumento proporcional en la probabilidad de acertar el ítem. En tal caso, se recurre a una función matemática que permita expresar la no linealidad de la relación y que además tenga en cuenta que la probabilidad de acierto del ítem está entre 0 y 1. Es así como se emplea una función logística[1] cuando existe una respuesta dicótoma en el ítem, lo que implica que hay una opción correcta y otras incorrectas, independientemente del número de opciones de respuesta del ítem. A esto se suman 3 parámetros de los ítems que son la dificultad del ítem *(b)*, la discriminación *(a)* y la probabilidad de acierto casual *(c)*.

El parámetro b hace referencia a la dificultad del ítem y representa la habilidad correspondiente a la probabilidad de contestar correctamente el ítem, siendo c el parámetro de acierto casual. La habilidad está en un rango entre $-\infty$ a ∞ pero se sitúa generalmente en un rango entre -4 y 4. Por definición, el parámetro de dificultad está en la misma escala de la habilidad, lo que garantiza una mayor interpretación de los resultados. Así, a medida que aumenta el valor del parámetro b el ítem se hace más difícil y cuando se reduce el valor de dicho parámetro, la dificultad del ítem disminuye.

El parámetro a corresponde a la discriminación y representa el cambio en la probabilidad de contestar correctamente el ítem conforme cambia el nivel de habilidad, en el punto correspondiente a la dificultad. El valor que toma este parámetro equivale a la pendiente de la recta tangente de la curva correspondiente a la dificultad en la curva característica del ítem, por lo cual una pendiente más pronunciada implica mayor discriminación. Lo anterior, dado que, a mayor pendiente, mayor es el cambio en la probabilidad de contestar correctamente, alrededor de la habilidad que corresponde a la dificultad del ítem.

Por último, el parámetro de acierto casual c hace referencia a la probabilidad de acertar el ítem al azar teniendo una baja habilidad. Dicho parámetro se conoce como asíntota inferior, pues es el valor asintótico que toma la probabilidad de acierto en el ítem en la medida en que el puntaje θ del ítem va siendo cada vez menor. En este caso, cuando el valor del parámetro c es pequeño, la probabilidad de acierto es muy reducida para los evaluados con baja habilidad, pero si el valor del parámetro c es alto, aquellos evaluados con baja habilidad tienen una probabilidad mayor de acertar el ítem al azar.

[1] La función logística es un refinamiento de la función exponencial, la cual se empieza a estabilizar después de un punto crítico.

Así, la ecuación de un modelo 3PL es de la forma

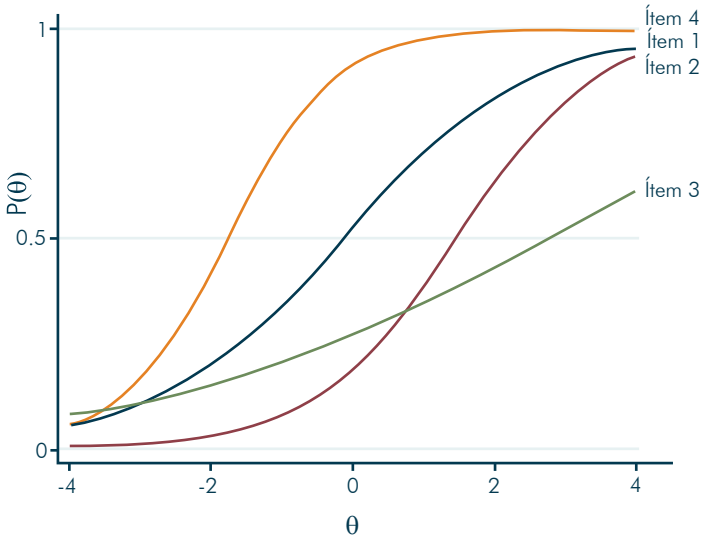
Ecuación 1.

$$P(X_j = 1 \mid \theta,a,b,c,) = P(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1+e^{1.7a_i(\theta - b_i)}}$$

Es decir, la probabilidad de acierto de un ítem X_j es condicional a su habilidad (θ) y a los parámetros del ítem (a,b,c).

En la figura 2 se evidencia cómo las características de diferentes ítems en términos de sus parámetros generan diferentes curvas características de ítems. En este caso, el ítem 2 tiene mayor dificultad que el ítem 4 y que el ítem 1, ya que la primera curva presenta un mayor desplazamiento a la derecha del eje x que las últimas dos. En cuanto al parámetro del azar, el ítem 3 tiene una probabilidad de acierto casual mayor a los demás ítems para aquellos evaluados de baja habilidad, lo cual se evidencia al comparar la probabilidad de acierto en los niveles más bajos de habilidad entre ítems, es decir cerca a θ=-4. De manera análoga, el ítem 2 tiene la probabilidad de acierto casual más baja con respecto a los otros ítems: para niveles de habilidad cercanos a -4 el ítem 2 tiene el nivel de probabilidad más cercano a cero. Alrededor de un θ=-2, se puede observar que el ítem 4 es el que mayor discriminación tiene, al tener la pendiente más pronunciada de las curvas características. Lo cual implica que el ítem 4 logra diferenciar muy bien entre evaluados con alta y baja habilidad en ese punto. Cabe señalar que la interpretación del parámetro de discriminación se realiza alrededor de la habilidad correspondiente a la dificultad.

Figura 2. Curvas características de los ítems con diferentes parámetros



Fuente: ICFES, 2017

El modelo 3PL es una generalización de los modelos de TRI para variables dicótomas. Cuando no se asume acierto casual (c=0), se convierte en un modelo logístico de 2 parámetros (2PL), y al asumir adicionalmente que la discriminación es constante entre ítems, se convierte en un modelo logístico de 1 parámetro (1PL). El modelo de Rasch puede ser visto como un caso particular del modelo 1PL, cuando se asume constante e igual a 1 el valor asociado a la discriminación de los ítems.

3. ¿Cómo se aplica?

Para el modelo de 3PL se parte de los siguientes supuestos[2]:

1. **Unidimensionalidad del trazo latente:** Se asume que los ítems están midiendo una única variable latente θ. Esto quiere decir que la prueba está midiendo una habilidad θ y que las variables que puedan afectar la respuesta al ítem son tratados como ruidos del ítem (o errores aleatorios).

2. **Independencia local entre los ítems:** La respuesta de cada ítem es independiente entre sí, controlando por θ. Se asume que los ítems no están correlacionados entre sí, así que la probabilidad de acertar un conjunto de ítems de la prueba corresponde a la multiplicación de la probabilidad de acertar cada ítem por separado.

3. **Invarianza de los ítems frente a los evaluados:** esta es la contribución teórica de medición y consiste en poder obtener mediciones que no varíen con respecto a los evaluados y a las pruebas empleadas.

La estimación de los tres parámetros se realiza para cada uno de los ítems. En general, se asume una distribución normal de θ dentro del proceso de estimación, esto es, que la distribución de las habilidades es simétrica alrededor de la media y que valores cercanos a la media son más probables que valores lejanos. Para lograr estimar la dificultad, la discriminación y la probabilidad de acierto casual de cada ítem, se maximiza la función de verosimilitud de los datos, la cual se compone por la multiplicación de las probabilidades de respuesta a cada ítem del test. Es decir, la función de verosimilitud es el producto del patrón de respuestas, y por tanto incluye tanto respuestas correctas P(θ) como respuestas incorrectas 1-P(θ).

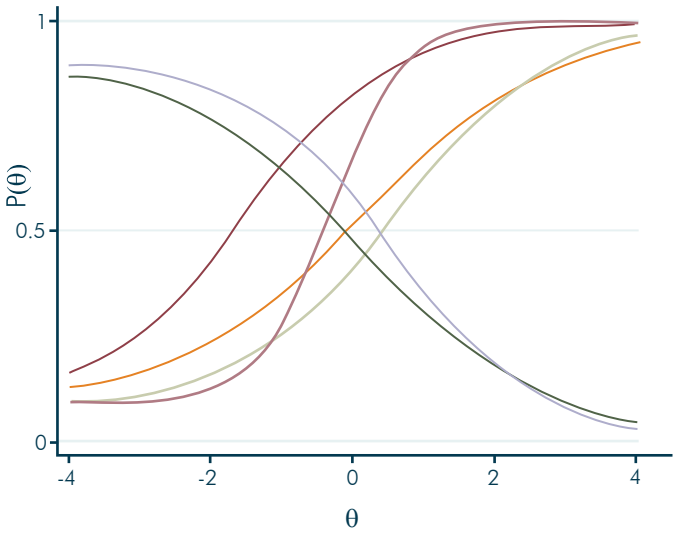
Una vez encontrados dichos parámetros, se calculan los nuevos valores de las habilidades y se repite este procedimiento de maximización hasta que el valor de cada parámetro no cambie sustancialmente. Si los parámetros de los ítems han sido estimados con otras pruebas, es decir los parámetros han sido calibrados previamente, la forma de calcular el puntaje de la prueba se realiza mediante la función de verosimilitud del modelo de 3PL (ecuación 1).

De manera ilustrativa, se presenta en la figura 3 un patrón de respuestas para 6 ítems que contiene 4 respuestas correctas y dos incorrectas. En este caso, las curvas de los ítems en forma de S invertida son las correspondientes a las respuestas incorrectas. Se puede apreciar que a medida que aumenta la habilidad (θ), aumenta la probabilidad de respuesta correcta de los 4 ítems y a la vez disminuye la probabilidad de las dos respuestas incorrectas. Así, la función de verosimilitud del test (que en este caso se compone de 6 ítems) corresponde a la multiplicación de las 6 respuestas a los ítems, bajo el supuesto de independencia local mencionado anteriormente. De esta manera, la estimación de la habilidad (θ) está condicionada al conocimiento de los valores de los parámetros.

En la figura 4 se presenta la función de verosimilitud de θ para el evaluado con dicha cadena de respuestas en el test de 6 ítems mencionado anteriormente. A partir de esa función, puede estimarse el valor $\hat{\theta}$ que maximiza la verosimilitud y que en este caso corresponde a un valor de 1. Este valor es el que con mayor verosimilitud tendría un evaluado que contestó los 6 ítems, generando la cadena de respuesta 111100, es decir, 4 respuestas correctas y 2 incorrectas, para los 6 ítems con las curvas presentadas en la figura 3. Nótese que si el valor estimado $\hat{\theta}$ hubiera sido -2, sería muy poco probable que el evaluado hubiera contestado correctamente los 4 ítems, dadas los valores de la probabilidad en ese punto, pero hubiera sido probable que hubiera contestado incorrectamente los otros dos ítems, es decir, no es muy verosímil que esa fuera su estimación de la habilidad. En el otro extremo, si la habilidad estimada $\hat{\theta}$ hubiera sido 3, es probable que el evaluado hubiera contestado correctamente los 4 ítems, pero poco probable que hubiera fallado en los otros 2, luego esta tampoco sería una estimación verosímil de su habilidad. El valor de 1, que corresponde a su estimación, es el que maximiza la probabilidad de haber tenido justamente el patrón de respuestas que tuvo el evaluado. Estimada la habilidad $\hat{\theta}$, se calcula la puntuación correspondiente, haciendo una transformación de la habilidad del evaluado en una escala definida.

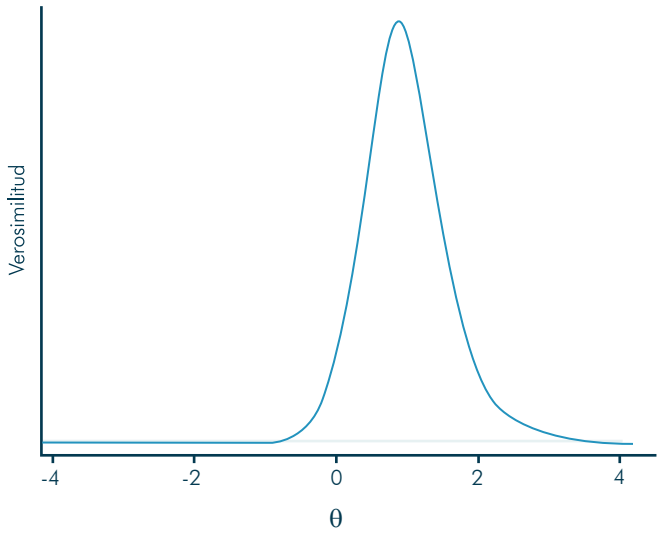
[2] La recopilación de los supuestos del modelo TRI 3PL se extrae de Muñiz (1997), Demars (2010) y Linden (2016).

Figura 3. Probabilidad de acierto de 6 ítems (111100)



Fuente: ICFES, 2017

Figura 4. Función de verosimilitud



Fuente: ICFES, 2017

4. Relevancia

La relevancia de usar un modelo de tres parámetros sobre otras especificaciones tiene que ver con la violación de la condición exigida sobre probabilidad cero de acierto al azar ($c=0$). Al emplear opciones de selección múltiple, resulta favorecedor emplear el modelo 3PL, generando mayor ajuste al ser más probable que existan elementos de azar relacionados e ítems con diferentes grados de discriminación sin afectar la precisión de las estimaciones (Muñiz, 1997). La mayor ventaja del modelo de 3PL es entonces la posibilidad de conocer más características de los ítems que componen el test. En el caso del Icfes, la creación de ítems contempla la evaluación de competencias en las áreas básicas del conocimiento, teniendo en cuenta diversos niveles de dificultad. De manera paralela a dicha construcción, se analiza el comportamiento de ítems y se toman en consideración los parámetros calibrados con el fin de mantener los estándares de comparabilidad. Gracias al armado actual de las pruebas Saber, estableciendo bloques con ciertas características (a los cuales llamamos bloques incompletos balanceados), es posible aumentar el número de ítems con diferentes grados de dificultad, discriminación y acierto casual y mantener la precisión en las estimaciones. De esta manera, modelar las pruebas Saber aplicadas por el Icfes empleando un modelo 3PL permite una mejor caracterización de los ítems, mayor precisión en las habilidades estimadas de los evaluados y mantener la comparabilidad entre periodos.

En próximos números de Saber en detalle, abordaremos justamente los temas relacionados con el armado de la prueba, la forma de garantizar la comparabilidad de los resultados, entre otros detalles técnicos de interés.

Bibliografía

Demars, C. (2010). Item Response Theory. Oxford University Press.
Linden, W. J. (2016). Handbook of item response theory. Boca Ratón: Taylor & Francis Group,LLC.
Muñiz, J. (1997). Introducción a la teoría de respuesta a los ítems. Madrid: Ediciones Pirámide.

SABER AL DETALLE

¿CÓMO SE GENERAN LOS PUNTAJES EN LAS PRUEBAS SABER DEL ICFES?

icfes mejor saber

GOBIERNO DE COLOMBIA

@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