

SERIE DE DOCUMENTOS TÉCNICOS / 2

APRENDER 2018

NOTAS

TÉCNICAS

SERIE DE DOCUMENTOS TÉCNICOS / 2

APRENDER 2018

NOTAS

TÉCNICAS

AUTORIDADES

Presidente de la Nación

Mauricio Macri

Jefe de Gabinete de Ministros

Marcos Peña

Ministro de Educación, Cultura, Ciencia y Tecnología de la Nación

Alejandro Finocchiario

Secretario de Gobierno de Cultura

Pablo Avelluto

Secretario de Gobierno de Ciencia, Tecnología e Innovación Productiva

Lino Barañao

Titular de la Unidad de Coordinación General del Ministerio de Educación, Cultura, Ciencia y Tecnología

Manuel Vidal

Secretaria de Evaluación Educativa

Elena Duro

Secretario de Gestión Educativa

Oscar Ghillione

Secretario de Políticas Universitarias

Pablo Domenichini

Secretaria de Innovación y Calidad Educativa

Mercedes Miguel

SECRETARÍA DE EVALUACIÓN EDUCATIVA

Secretaría de Evaluación Educativa

Elena Duro

Directora Nacional de la Evaluación de la Calidad y Equidad Educativa

María Cortelezzi

Directora de Gestión

Paula Camarda

Coordinador de Información y Comunicación

Gustavo Streger

Coordinador de Metodología

Augusto E. Hoszowski

Coordinadora de Evaluaciones del Nivel Superior

María Florencia Buide

EQUIPO DE LA SECRETARÍA DE EVALUACIÓN EDUCATIVA

5

Adriana Luján
Anabella Julieta Díaz
Andrés Nussbaum
Antonella Guerrero
Antonella Misenti
Belén Santillán
Betina Presman
Camila Mansilla
Cármén de La Linde
Cecilia Beloqui
Cecilia Toledo
Christian Rodriguez
Claudia Comparatore
Clarisa Cortés
Diego Hernán Nigro
Elda Salinas
Emiliana García
Florencia Carballido
Francisco Forlizzi
Gissella Mernies Ivanoff

Gloria Landa
Graciela Fernandez Trelles
Hernán Lakner
Inés Marie Laxague
Jezabel Danon
Jorge Kancepolsky
Juan José Gorostiague
Juliana Córdoba
Julieta Jimenez
Karen Martínez
Laura Golszczyznsky
Laura Melchiorre
Liliana Bronzina
Lucía Iannaccone
Malena Otaduy
María Aranguren
María Asunción Fragni
María Clara Radunsky
María Julieta
Mesón Aragón

María Elena Brenlla
María Soledad Pino
María Victoria Beherengaray
Calvo
Mariana Mora
Mariela Ceschel
Marilina Zayas
Martina García Reynoso
Melody Guevara
Mirta Beatriz Colle
Montserrat Salto Lastra
Natalia Pereyra Ernst
Patricia Graciela Oppel
Quimey Lassus Frusso
Rubén Marcial

Asesores

Rafael Mallo
Romina Ángel
Rosalía Cortés

Modelo de teoría de respuesta al ítem (TRI)

Formas en la prueba Aprender

En las evaluaciones educativas estandarizadas (a diferencia de las evaluaciones de certificación) es usual que los alumnos no respondan todos al mismo conjunto de preguntas. El motivo es que si todos los alumnos tuvieran que responder un conjunto de preguntas que abarquen todos los temas de una disciplina, el cuestionario sería demasiado extenso. Se opta entonces por dividir el conjunto de preguntas en *formas* o *modelos*. La práctica usual es dividir el conjunto de preguntas en bloques de igual cantidad de preguntas, en lo posible de dificultades medias similares y abarcativos del conjunto de temas focalizados. Con estos bloques se construyen los modelos, variando los bloques que los componen o el orden en que estos aparecen.

En Aprender 2016, 2017, 2018 el total de preguntas es 72 para cada disciplina y año. Cada bloque contiene 12 ítems y hay un total de 6 modelos, en cada año y disciplina.

En Aprender el diseño de estos seis modelos es el siguiente:

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6
1° bloque	Bloque 1	Bloque 2	Bloque 3	Bloque 4	Bloque 5	Bloque 6
2° bloque	Bloque 4	Bloque 5	Bloque 6	Bloque 2	Bloque 3	Bloque 1

Vemos que cada bloque aparece en primer lugar en la misma cantidad de modelos en que aparece en segundo lugar. Cada bloque aparece en la misma cantidad de modelos.

Este esquema de diseño de los modelos se denomina *balanceado incompleto*: cada bloque aparece en cierta posición en el mismo número de modelos y cada par de bloques aparece en algún modelo (balanceo) pero no todos los bloques aparecen en los modelos (*incompleto*).

Los especialistas construyen los modelos de manera tal que sean lo más parecidos entre sí posible, tanto en lo que refiere a nivel de dificultad, como en las temáticas abarcadas y en la cantidad de ítems por temática.

En cada sección evaluada se distribuyen los 6 modelos en forma aleatoria, cuidando de aplicar, en la medida de lo posible, la misma cantidad de modelos en cada aula. En Argentina se utiliza más comúnmente la palabra *modelo* que *forma*.

La Teoría de Respuesta al Ítem

Las preguntas (ítems) de las evaluaciones Aprender 2018 (que son preguntas de respuesta múltiple, con cuatro opciones, una sola de ellas correcta) se analizaron en base al modelo de la Teoría de Respuesta al Ítem (TRI), en la que se fundamenta gran parte de las evaluaciones estandarizadas internacionales, así como las evaluaciones del Operativo Nacional de Evaluación (ONE) entre 2005 y 2013, Aprender 2016 y Aprender 2017..

Los supuestos básicos del TRI son:

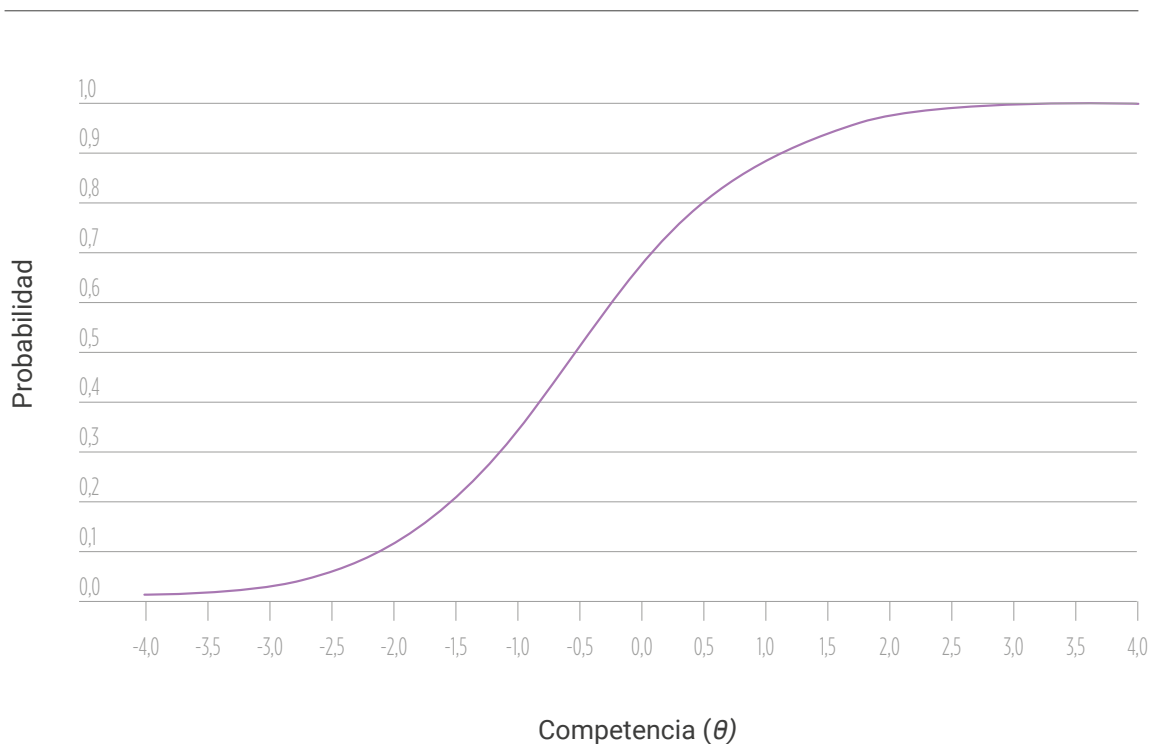
- a) Para cada disciplina evaluada en cierto año, cada alumno posee una *habilidad, rasgo latente* (no observado) o *competencia*, que puede asociarse con un número real θ . Hay que señalar que, de uso común en la psicometría, los términos *rasgo latente, habilidad o competencia* no implican ni hacen referencia a alguna característica innata de los sujetos a los cuales se les aplica el test.
- b) La probabilidad de responder correctamente determinado ítem (pregunta) es una función creciente y continua del valor θ .
- c) Unidimensionalidad: una sola *habilidad* (medida con un número real θ) explica la probabilidad de respuesta correcta a cada ítem en cierta disciplina y año.
- d) Independencia local: dado un cierto nivel de competencia θ , el responder correctamente dos ítems distintos i y j son eventos *independientes*.

Aunque estos son los supuestos usuales, hay modelos TRI donde se postula *multidimensionalidad*, o donde cada alumno no tiene un valor constante asociado θ , sino que θ es una variable aleatoria, reportándose algunos valores de esta variable (ver documentos técnicos PISA para esto).

Como en todo modelo, estos supuestos nunca se cumplen exactamente, pero la experiencia en las evaluaciones educativas muestra que el modelo TRI se ajusta razonablemente bien. Sin embargo, es necesario que todos los pasos de la prueba sean diseñados y ejecutados con cuidado, desde su marco teórico, el diseño de los ítems, hasta su aplicación.

Para modelizar la función de respuesta del punto b), se elige habitualmente una función que esté comprendida en el intervalo (0;1). Se parte del supuesto de que por más elevado que sea el nivel de competencia, la probabilidad de respuesta nunca será igual a uno. Y simétricamente, el supuesto es que aunque el nivel de competencia sea muy bajo, la probabilidad de respuesta nunca será igual a cero. Intuitivamente, la forma de la función de respuesta puede ser representada mediante el gráfico siguiente:

Función de respuesta



Hay modelos donde la curva de respuesta no tiene como asíntota horizontal inferior al eje X, sino una recta horizontal con $Y > 0$. Lo que se interpreta como que por más pequeño que sea el parámetro θ , la probabilidad de respuesta se mantiene siempre mayor al valor de la asíntota. Una justificación a esto es que en las pruebas de opción múltiple, el azar permite marcar la respuesta correcta, independientemente de la dificultad del ítem. Aquí también aparece el supuesto de que la prueba es respondida de 'buena fe', y no en forma azarosa, por ejemplo.

Una de las opciones más usuales es la función logística. Según este modelo, la probabilidad que tiene un alumno con cierto nivel de competencia θ de responder correctamente al ítem i es

$$Prob(I_i = 1) = \frac{\exp(a_i \cdot (\theta - b_i))}{1 + \exp(a_i \cdot (\theta - b_i))}$$

donde a_i es la discriminación del ítem i y b_i es la dificultad del ítem i y θ la competencia del alumno en cuestión.

La expresión anterior tiene estas propiedades, suponiendo que la discriminación a_i es positiva:

- La función es estrictamente creciente, como función de θ : A mayor habilidad, mayor probabilidad de responder correctamente el ítem i .
- El límite de la función, cuando θ tiende a ∞ , es 1
- El límite de la función, cuando θ tiende a $-\infty$ es 0
- Si $\theta = b_i$, la probabilidad de responder correctamente el ítem i es 0.5

Si se postula que todos los a_i son iguales se habla de un modelo logístico a un parámetro (solo el nivel de dificultad). Aprender se modeliza con un modelo logístico a dos parámetros, el más usual actualmente. Este modelo permite no solo ajustar el nivel de dificultad de los ítems sino también su capacidad de discriminación (parámetro a_i).

En principio si al estimar el modelo encontramos que la discriminación es negativa o cercana a cero, el ítem debe descartarse durante el proceso de análisis.

Dificultades de los ítems

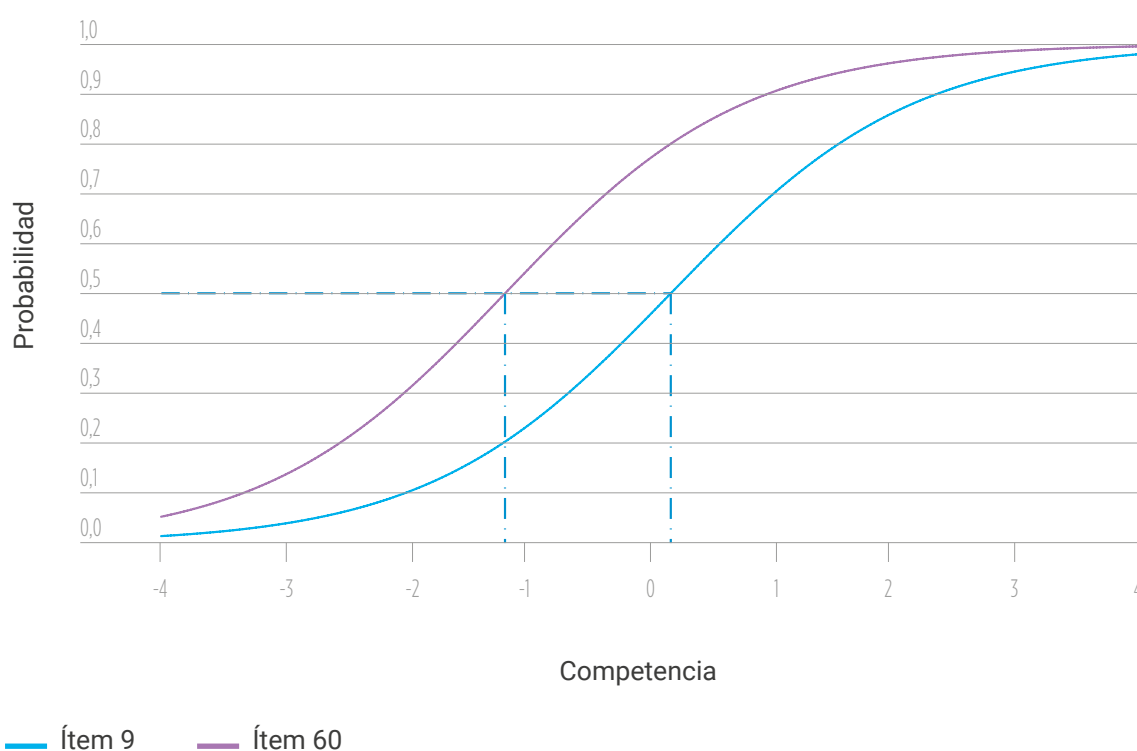
En el TRI cada ítem (pregunta) tiene asociado un parámetro, tradicionalmente denotado con la letra b , denominado dificultad. Este parámetro está en la misma escala que las competencias, por lo que teóricamente puede tomar cualquier valor real. En la práctica, si las competencias se ajustan a una escala de media cero y desvío standard 1, los ítems estarán en general en el intervalo $(-3 ; 3)$. Intuitivamente, a mayor valor de b , menor será la proporción de alumnos que responden correctamente el ítem, y si un ítem tiene $b=0$, el 50% lo responderá correctamente con probabilidad=0.5 (si las competencias siguen una distribución normal de media cero).

El parámetro b indica el corrimiento de la curva respecto al cero.

En el modelo TRI usual, la curva logística se corre a la derecha o la izquierda según una mayor o menor dificultad, como se ejemplifica en el siguiente gráfico:

9

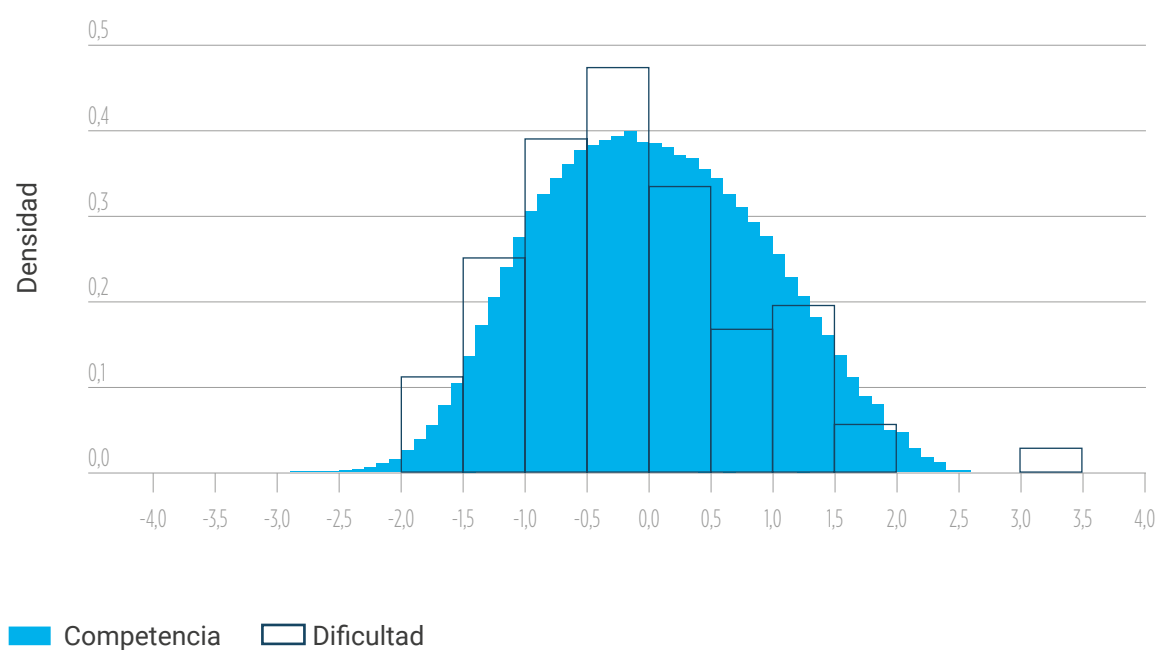
Curva característica de los ítems 9 y 60, 6º año en Lengua



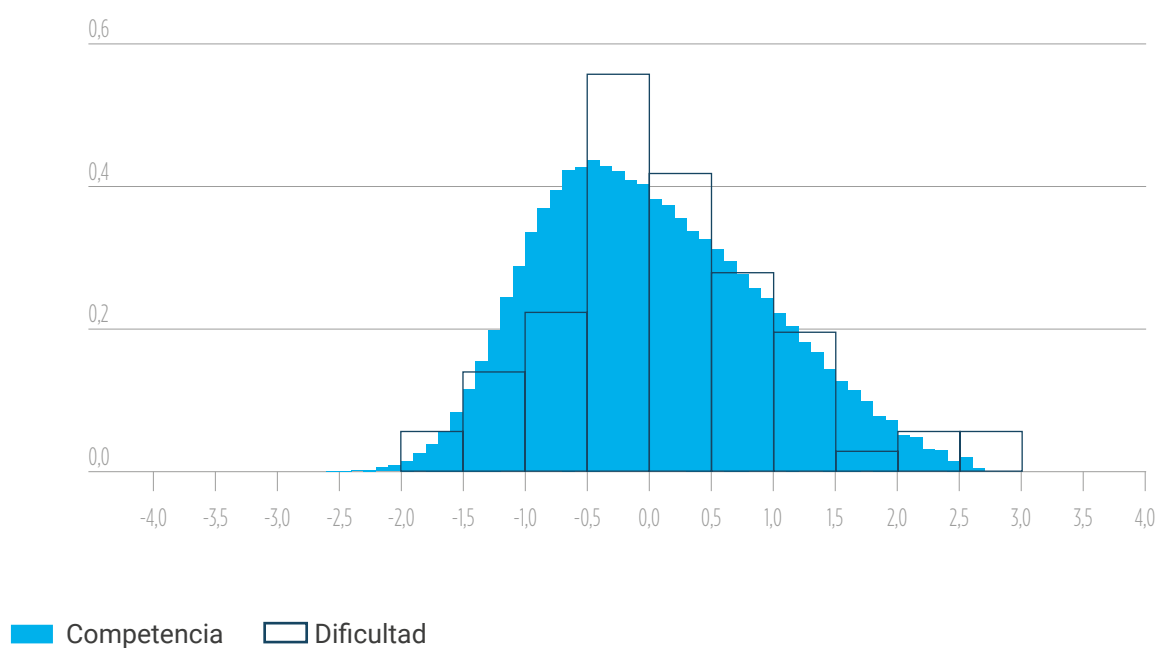
Una de las recomendaciones para la construcción de pruebas estandarizadas es que las dificultades deben barrer todo el espectro de competencias para, de esta forma, poder estimar de la manera más ajustada posible la competencia de cada alumno.

Si escalamos las competencias de los alumnos en el intervalo $(-3 ; 3)$, las dificultades de los 72 ítems de Aprender 2018 resultaron ser:

Distribución de competencias y dificultades, 6º año en Matemática



Distribución de competencias y dificultades, 6º año en Lengua



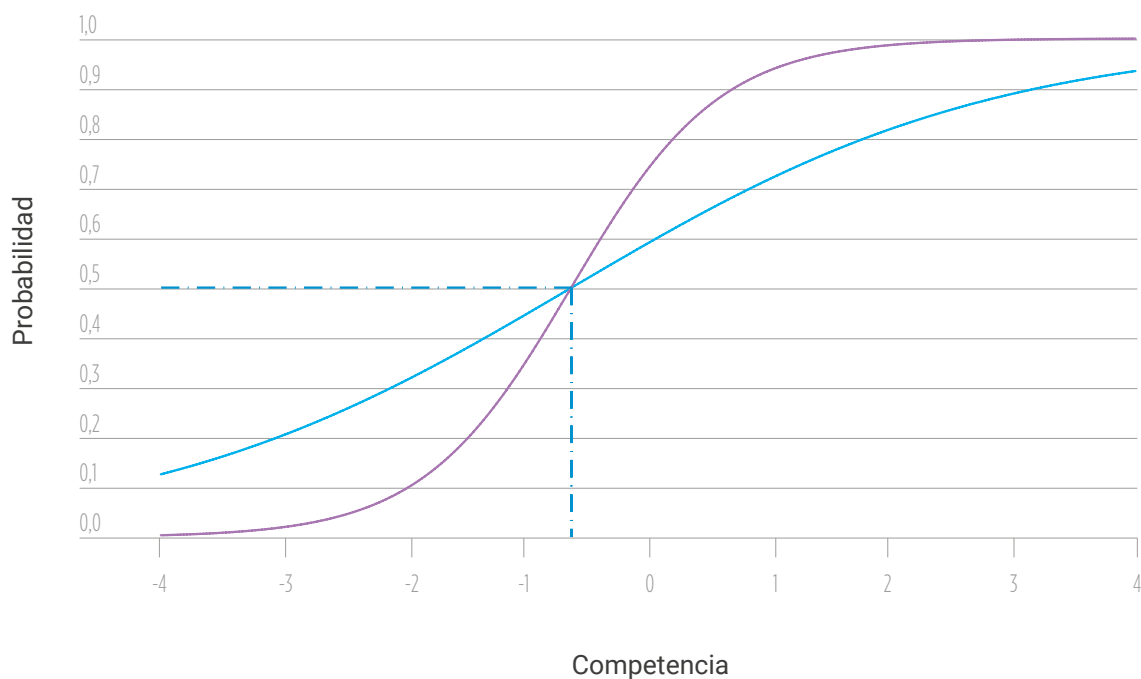
Vemos que las dificultades presentaron una buena dispersión dentro del rango de las competencias.



Nivel de discriminación de los ítems

En el modelo a dos parámetros (aplicado en Aprender), el segundo parámetro es la discriminación (denotado en general con la letra a), que determina la inclinación de la curva en el entorno de b . Ejemplificamos dos ítems con dificultades similares pero diferente discriminación.

Curva característica de los ítems 23 y 28, 6º año en Lengua

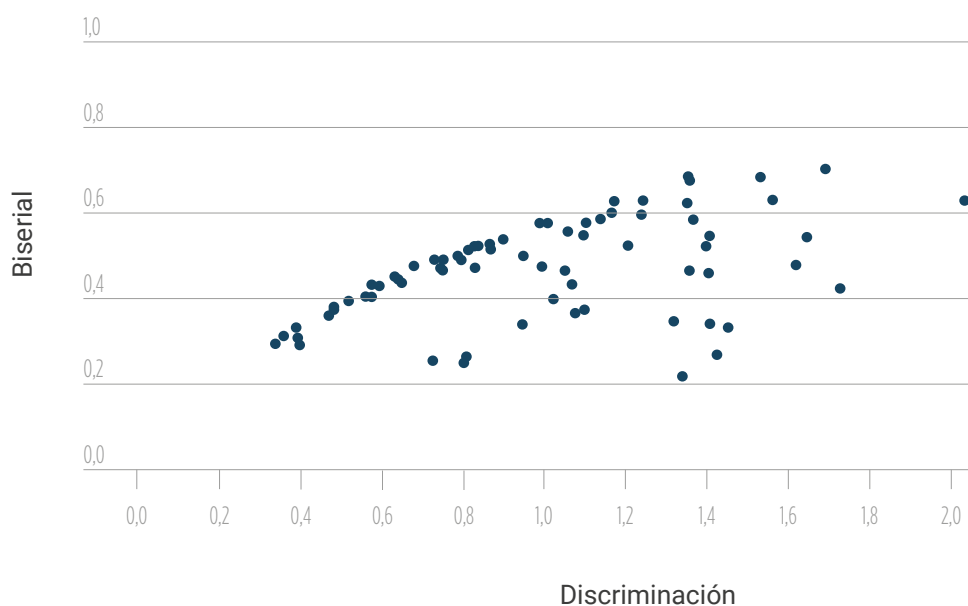


La discriminación de los ítems (supuesto un modelo logístico a dos parámetros) deber ser como mínimo positiva (a mayor nivel de competencia, mayor probabilidad de respuesta correcta).

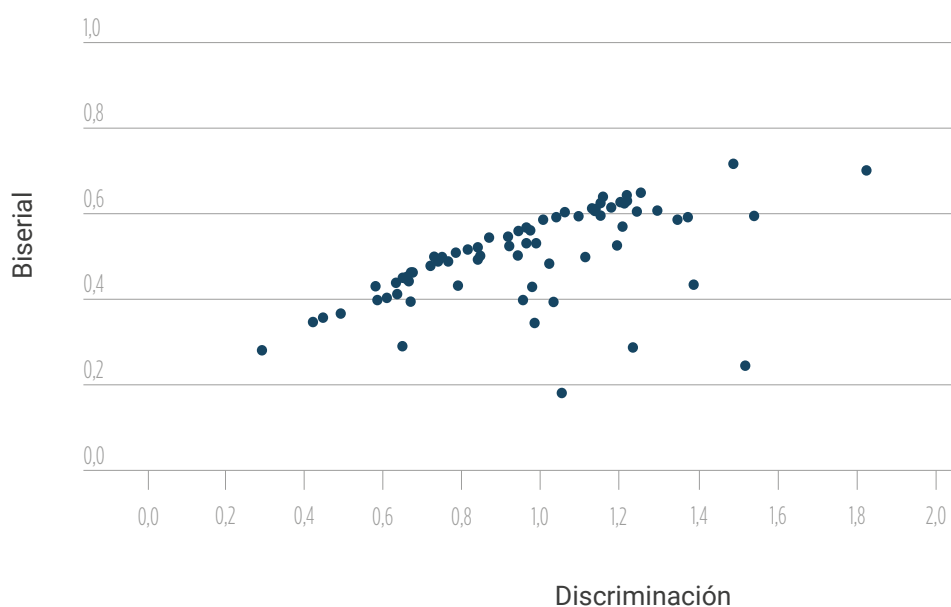
El nombre *discriminación* se debe a que cuanto mayor sea a (supuestamente positivo), mayor es el poder del ítem para *discriminar* a los alumnos con una competencia mayor que b de los alumnos con una competencia menor que b .

Se sugiere también que las discriminaciones sean mayores a 1.15, en el caso de estar las competencias normalizadas con media cero y desvío 1. En Aprender 2018 6º año Matemática la distribución de las discriminaciones fue:

Correlación biserial - Discriminación TRI, 6º año en Lengua



Correlación biserial - Discriminación TRI, 6º año en Matemática



lo que muestra un buen funcionamiento de los 72 ítems. Una recomendación usual es que un ítem, bajo el modelo TRI, posee una discriminación aceptable si esta es aproximadamente superior a 0.15. En la práctica otros factores pueden hacer recomendable conservar un ítem que tiene una discriminación positiva inferior a 0.15: por ejemplo, la importancia de la temática, su correlación ítem-test, patrones de respuesta al ítem, etc. El equipo pedagógico de la SEE realiza un análisis global de estos factores antes de decidir eliminar un ítem de la prueba.

Ajuste del modelo

En los distintos cuestionarios donde se ha aplicado, se verificó que los supuestos del TRI se cumplen en forma satisfactoria. Para garantizar la calidad de la información producida, Aprender realiza, luego de cada prueba, un "testeo" del funcionamiento del modelo y del funcionamiento de los ítems (e.g. que las discriminaciones sean positivas). El ajuste del modelo (nivel de concordancia del modelo supuesto con la información relevada, i.g. las respuestas de los alumnos a los ítems) se puede hacer por media de tests.

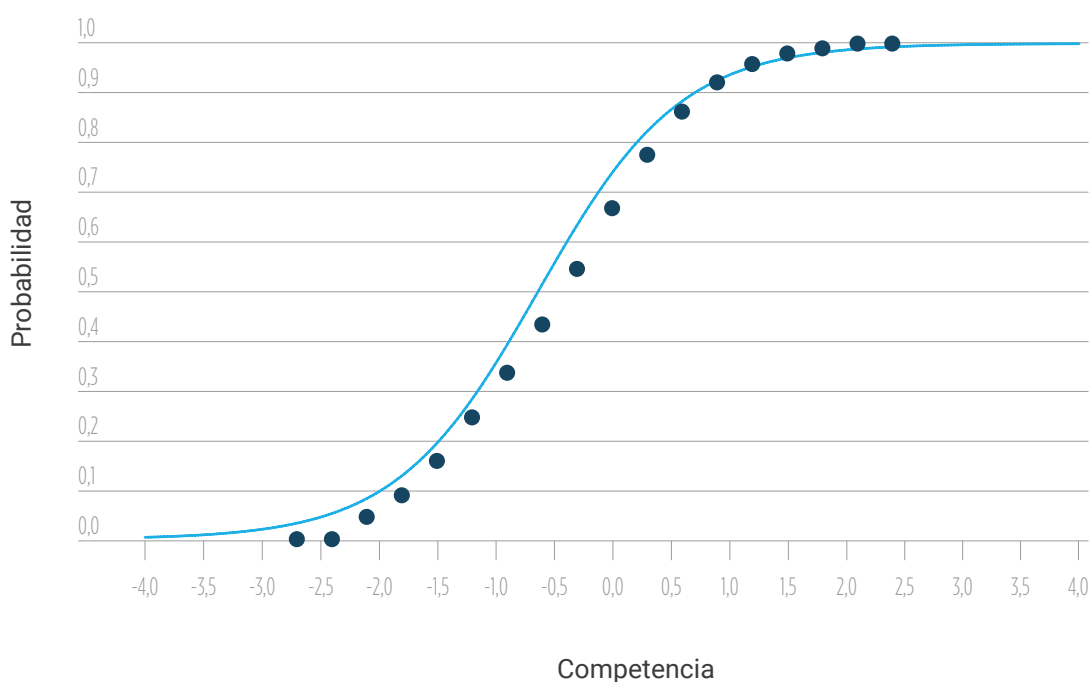
a. Ajuste de la función de respuesta

Lo primero que se testea es si la función de respuesta supuesta (modelo logístico con dos parámetros) se corresponde con el patrón de respuestas observado.

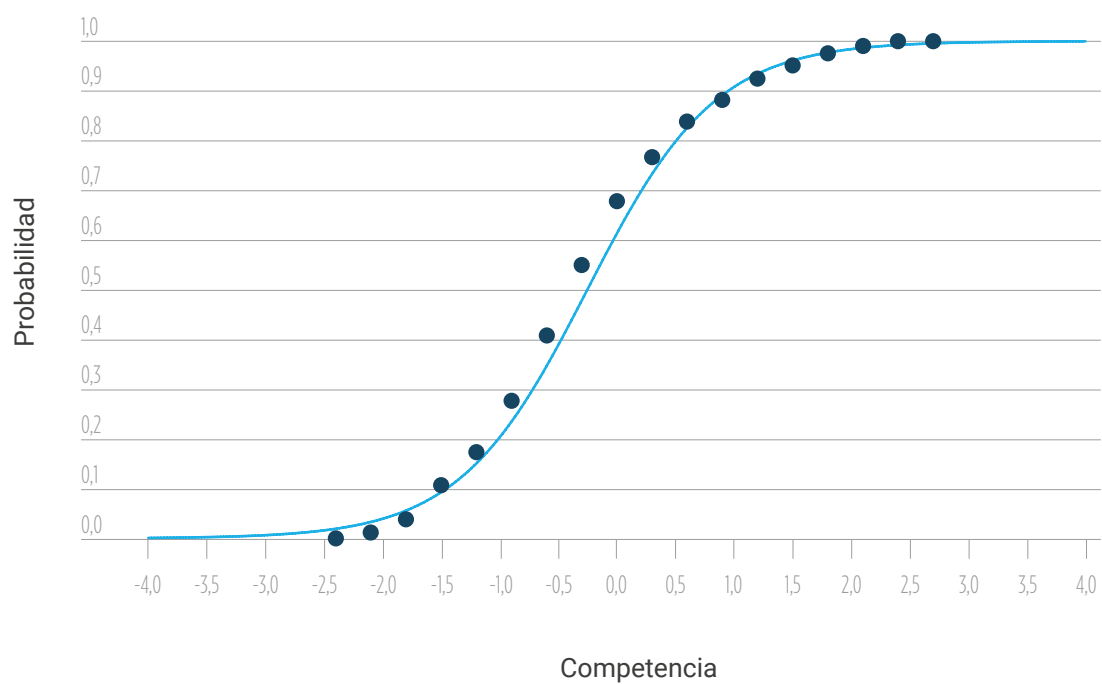
Para testear esto lo ideal sería, para cierto valor de θ , administrar el mismo ítem n veces a un alumno con nivel de competencia θ y verificar que la proporción de respuestas correctas es la predicha, o se encuentra dentro de cierto intervalo de confianza. Como esto no es posible (por el efecto aprendizaje entre otros factores) se hace lo siguiente: dividir el rango de valores θ estimados en k tramos y en cada tramo hallar la proporción de respuestas correctas. Esta proporción será la proporción muestral de aciertos. Si los intervalos son relativamente angostos, las probabilidades teóricas de acierto no deberían diferir mucho en cada intervalo, por lo que podemos suponer que la proporción hallada debería representar a la teórica en el punto medio del intervalo. Con estas proporciones muestrales, se puede aplicar un test de ajuste, chi-cuadrado, entre otros.

Como ejemplo de esto se grafican las funciones de respuesta observadas y empíricas de tres ítems.

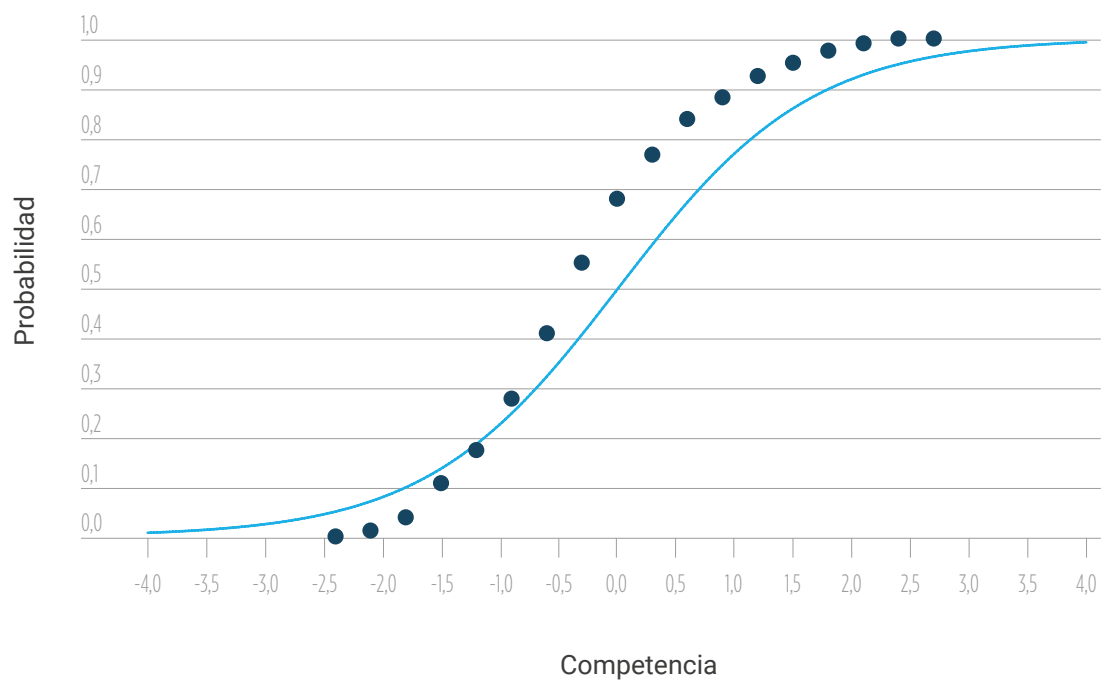
Curva característica del ítem 28, 6º año en Lengua



Curva característica del ítem 31, 6º grado en Matemática



Curva característica del ítem 57, 6º año en Matemática



b. Unidimensionalidad

Siendo uno de los supuestos básicos del modelo postulado, su testeo se realiza por medio de un análisis factorial del conjunto de variables dicotómicas Verdadero/Falso de las respuestas. Al ser variables dicotómicas, el análisis factorial se realiza sobre la matriz de correlaciones tetracóricas. Si el supuesto de unidimensionalidad se cumple es de esperar que el primer autovalor sea mucho mayor que los demás. En Aprender las estimaciones de estas correlaciones se hicieron con el software Stata. Se presentan a continuación los gráficos de los autovalores estimados para algunos modelos de Aprender 2018. Para evitar la presencia de los valores faltantes se procesó cada modelo en forma independiente.

Gráfico de autovalores según correlación tetracórica, 6° año en Lengua (Modelo 1)

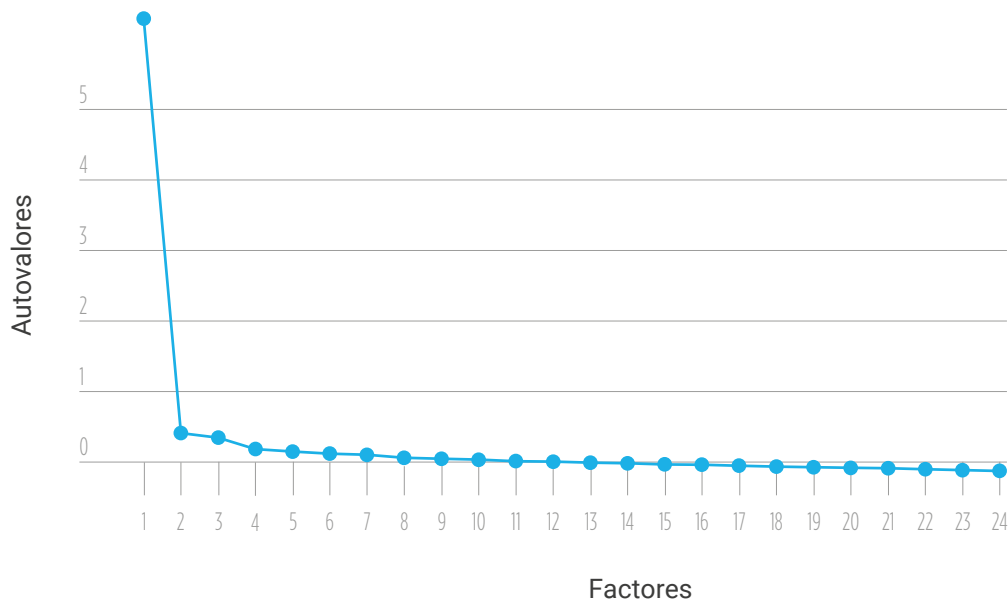
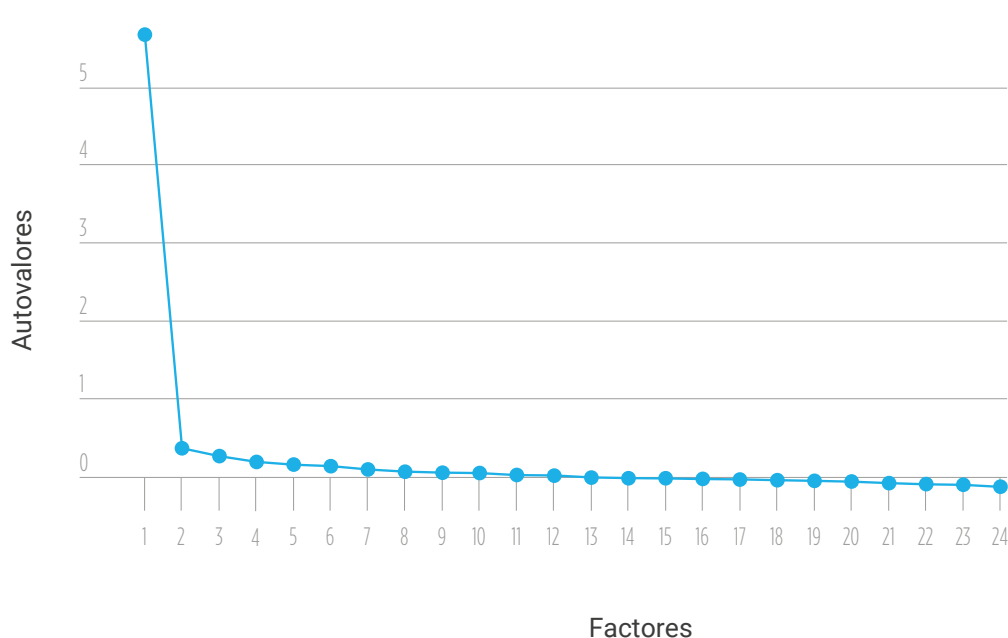


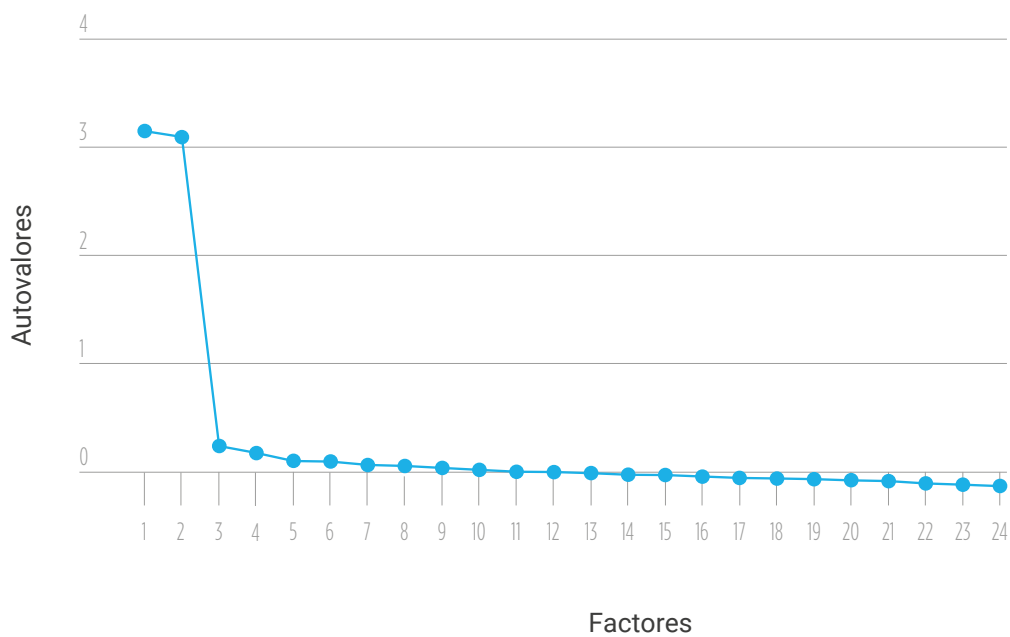
Gráfico de autovalores según correlación tetracórica, 6° año en Matemática (Modelo 2)



Se verifica que en todos los casos el primer factor retiene la mayor parte de la varianza.

Para ejemplificar un caso de no unidimensionalidad se realizó este experimento: Se agregaron las respuestas de Lengua y Matemática, modelo 1. Dejando solo los alumnos que respondieron a este modelo. Y se repitió el cálculo de los factores y su correspondientes autovalores. El resultado muestra la existencia de dos factores, claramente con autovalores de orden similar.

Gráfico de autovalores según correlación tetracórica, 6° año en Lengua y Matemática (Modelo 1)

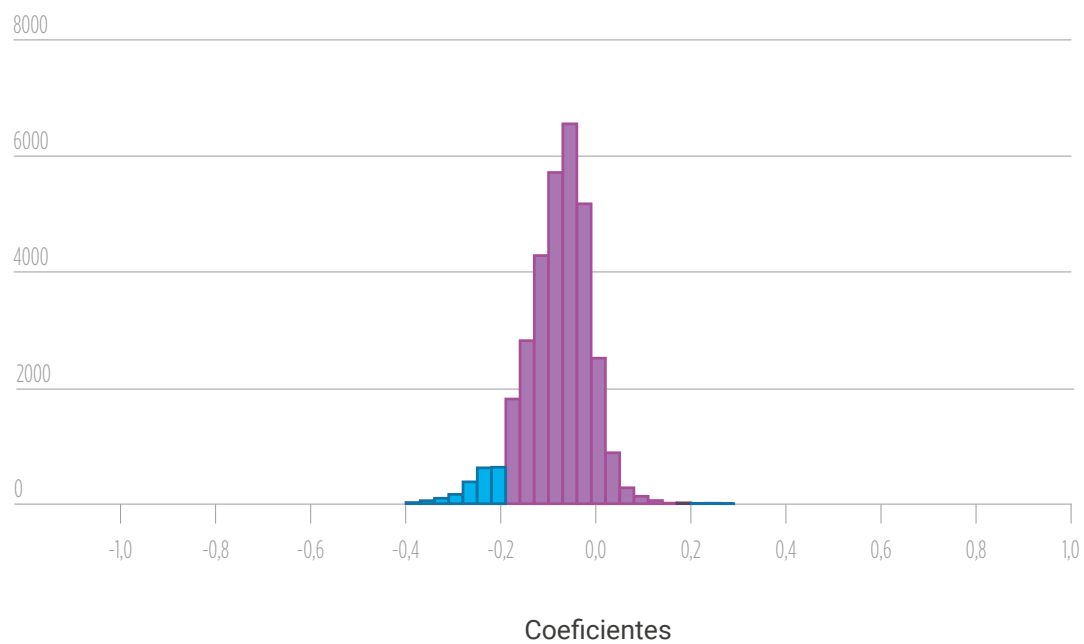


c. Independencia local

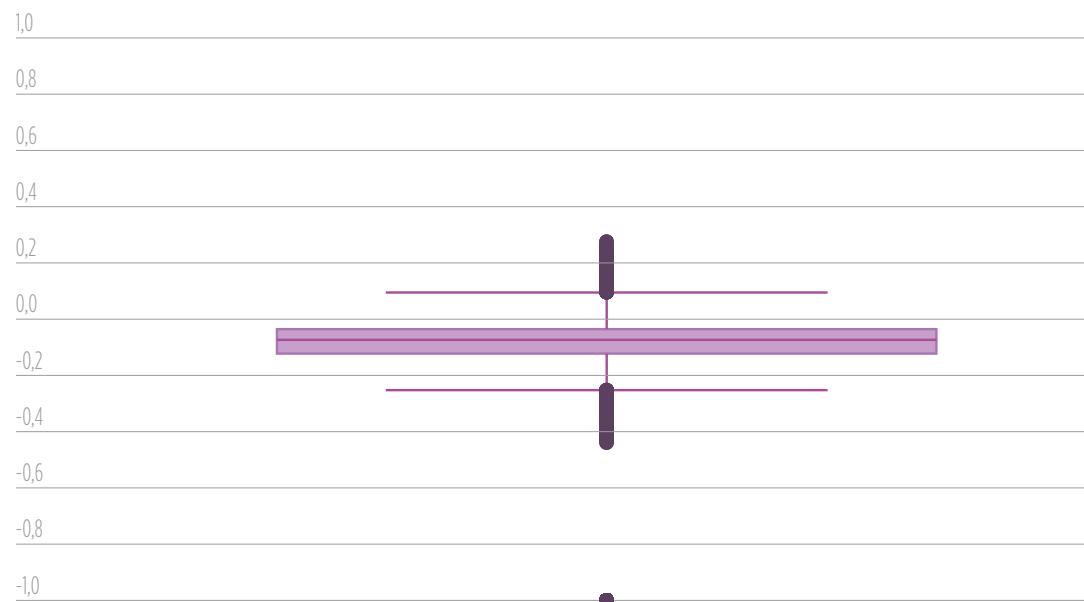
En el modelo TRI la independencia local juega un papel importante, principalmente a la hora de realizar las estimaciones de los parámetros. Si esta hipótesis se cumple, a niveles similares de competencia (valores de θ), las variables indicadoras de respuesta correcta de dos ítems deben ser variables aleatorias independientes y por lo tanto sus correlaciones deben tener valores cercanos a cero. Una primera verificación puede ser, entonces, que las correlaciones estén cercanas a cero. Una alternativa es proceder en forma similar a la de unidimensionalidad. Se divide el rango de variación de las habilidades estimadas θ en k tramos, y en cada tramo se estiman las correlaciones tetracóricas entre los ítems, dentro de cierto modelo M (si hay r ítems en este modelo, habrá $r(r-1)/2$ correlaciones). Se obtiene entonces un conjunto de correlaciones, cada una correspondiente a un tramo y a un par de ítems. Este conjunto de correlaciones se grafica con un histograma, para verificar si su distribución se corresponde con la hipótesis de que son todas, o casi todas cercanas a cero (al ser estimaciones a partir de un conjunto de respuestas de elección múltiple habrá necesariamente algunas estimaciones alejadas del cero). El resultado de este análisis para 6º año Lengua y Matemática 2018, se presenta en el siguiente gráfico.



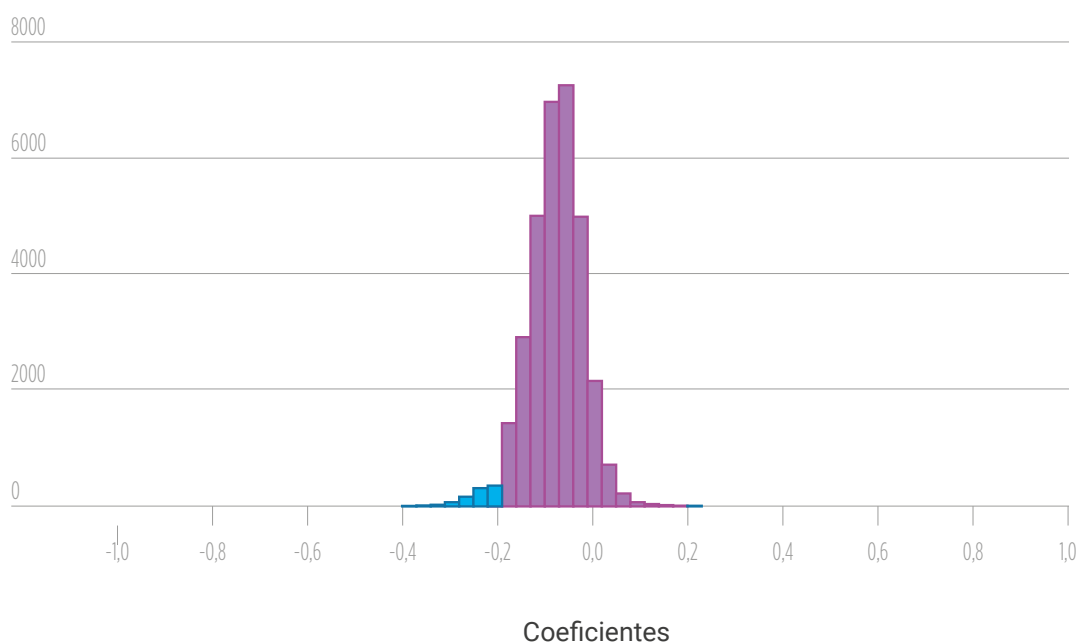
Correlación tetracórica inter ítems, 6º año en Lengua



Distribución de las correlaciones tetracóricas inter ítems, 6º año en Lengua



Correlación tetracórica inter ítem, 6º año en Matemática



Distribución de las correlaciones tetracóricas inter ítems, 6º año en Matemática



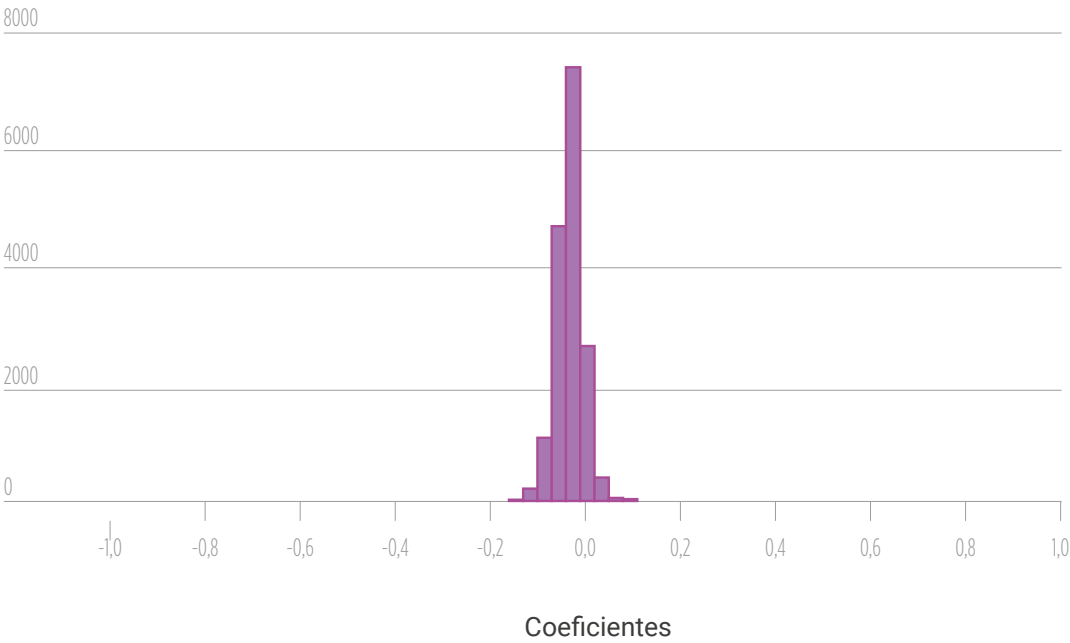
Un análisis más ajustado podría implicar el estudio de la correlación entre un par de ítems distintos mediante el estadístico Q3. Este estadístico se define como el coeficiente de correlación entre las variables:

$$\begin{aligned} r_i &= c_i - Prob(I_i = 1|\theta) \\ r_j &= c_j - Prob(I_j = 1|\theta) \end{aligned}$$

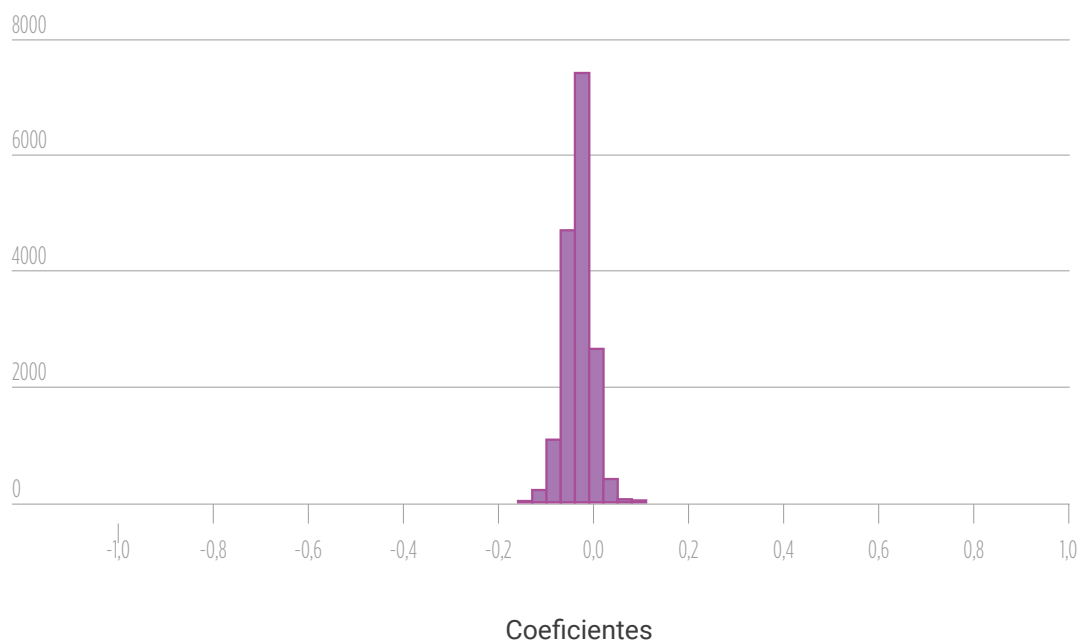
siendo c_i la variable indicadora de si la respuesta al ítem i fue correcta, y $Prob(I_i=1|\theta)$ la probabilidad de responder correctamente al ítem i , con cierta competencia θ . Al tomar los residuos se elimina el efecto de calcular la correlación con unidades de distinta habilidad.

Si calculamos para cada par i,j este coeficiente podemos graficar el histograma de estos valores:

Coeficientes de correlación inter ítems Q3, 6º año en Matemática



Coeficientes de correlación inter ítems Q3, 6º año en Lengua



22

Aprender 2018

Verificamos que los valores de Q3 se encuentran en el intervalos (-0.2 ; 0.2), que es el intervalo que se considera aceptable para considerar cumplida la hipótesis de independencia local.

La escala de puntajes en el modelo TRI

La escala TRI es arbitraria, se entiende por esto que no hay un cero 'natural' ni una unidad de medida natural, es decir que infinitas escalas son posibles. En Aprender se escalaron los valores θ mediante una transformación lineal (multiplicando por un factor y sumando una constante) para que tenga media 500 y desvío standard 100, tomando Aprender 2016 como punto de referencia.

Esta escala puede ser transformada no solo en forma lineal. Una transformación no lineal usual es la escala True Score.

Se le asigna a cada alumno con habilidad estimada θ el número esperado de respuestas correctas:

$$TS(\theta) = \sum_{i=1}^k P(I_i = 1|\theta)$$

Esta escala tiene la ventaja que su rango es el intervalo (0,k), donde k es el número de ítems en el modelo del alumno, lo que facilita su interpretación.

Esta escala está correlacionada con la original de media 500 y desvío standard 100, como lo muestra el siguiente scatter.

Puntaje Theta según Score,6º año en Lengua

