INTRODUCCIÓN A LA TEORÍA DE RESPUESTAS AL ÍTEM

TRI

Santo Domingo, República Dominicana
2014

Ministerio de Educación (MINERD)
Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa (IDEICE)

Investigador Principal:
Dr. Héctor Valdés Veloz

Diagramación y diseño:
Yeimy Olivier
Natasha Mercedes

Centro de Documentación:
Ing. Dilcia Armesto

ISBN: 978-9945-8859-8-9

Santo Domingo, República Dominicana.
2014

MINERD
Ministerio de Educación

AUTORIDADES

Lic. Danilo Medina Sánchez
Presidente de la República

Dra. Margarita Cedeño
Vicepresidenta de la República

Lic. Carlos Amarante Baret
Ministro de Educación

Lic. Luis Enrique Matos de la Rosa
Viceministros de Educación
Encargado de Asuntos Técnicos Pedagógicos

Dr. Jorge Adarberto Martínez Reyes
Viceministros de Educación
Encargado de Supervisión, Evaluación y Control de la Calidad Educativa

Ing. Víctor Ricardo Sánchez Jáquez
Viceministros de Educación
Encargado de la Oficina de Planificación Educativa, OPE

Licda. Vivian Elizabeth Báez Báez
Dirección General de Recursos Humanos

DIRECTORES EJECUTIVOS Y DIRECTORES GENERALES

Denia Burgos, M.Ed
Instituto Nacional de Formación y Capacitación del Magisterio, INAFOCAM

Dr. Julio Leonardo Valeirón Ureña
Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa, IDEICE

Lic. Rosa María Kasse Soto
Oficina de Cooperación Internacional del Ministerio de Educación

Dra. Ancell Scheker Mendoza
Dirección General de Evaluación de la Calidad Educativa

María Esperanza Ayala de la Cruz M.Ed
Dirección General de Supervisión Educativa

Dra. Carmen Margarita Sánchez Ramos
Dirección de Currículo

INTRODUCCION A LA TEORÍA DE RESPUESTA AL ÍTEM

Introducción

La teoría clásica del test (TCT) fue desarrollada durante los años veinte del siglo pasado.

En esta teoría el resultado de la medición de una variable dependía del test utilizado, lo que determinaba la existencia de serios problemas para tratar de establecer la equivalencia entre las puntuaciones de dos tests distintos que medían una misma variable, con lo cual era poco probable poder comparar los resultados de dos poblaciones examinadas con dos tests diferentes que trataban de medir sus rendimientos en aritmética, por ejemplo.

De manera que en la TCT la medida de una variable es inseparable del instrumento utilizado para medirla Esto constituye una seria limitación de la referida teoría, pues de manera operativa se define la variable por el instrumento que se utiliza para medirla.

Ya en el año 1928 Thurstone sentenció con toda claridad: "…un instrumento de medida no debe venir afectado por los objetos medidos… sus mediciones deben ser independientes de los objetos medidos". [1]

La limitación antes referida puede comprenderse con mayor claridad en el ejemplo siguiente:

Supongamos que el profesor de Matemática de Julio le aplica una prueba elaborada bajo los supuestos de la Teoría Clásica del Test para conocer su rendimiento académico en esa asignatura.

Semanas después, al profesor de Julio se le extravía el examen aplicado y entonces lo somete a una nueva prueba elaborada siguiendo la misma teoría. Horas más tarde aparece el primer examen aplicado. Al calificar ambas pruebas los resultados son bastante diferentes. El profesor se pregunta entonces: ¿cuál es el rendimiento académico de Julio en Matemática?

Por otra parte, en la T.C.T las propiedades del instrumento de medida, o sea de los ítems y del test, están en función de los sujetos a los que se les aplica.

Esto significa que, por ejemplo, el índice de dificultad de un ítem dependerá del nivel de competencia o de habilidad que tenga el grupo de sujetos que lo responde.

Las dos limitaciones de la TCT antes descritas sintéticamente demuestran que la misma estaba encerrada en una importante incongruencia teórica: la medición depende del instrumento utilizado y las propiedades de esta están determinadas por las características o nivel de habilidad de los sujetos que lo responden.

Para dar solución a estas limitaciones se desarrolló la Teoría de Respuesta al Item (TRI). Su nombre proviene del hecho de que su enfoque se basa en las propiedades de los ítems más que en las del test en sentido global.

1 Thurstone, L.L. Attitudes con su measured. American Journal of Sociology 1928, pág. 547.

Como Lord (1980) aseguró, la TRI no contradice ni las asunciones ni las conclusiones fundamentales de la Teoría Clásica de los Tests, sino que hace asunciones adicionales que permiten responder cuestiones esenciales que la TCT no podía.

Mientras que los conceptos básicos de la TRI eran, y son, sencillos, la matemática que la fundamenta era de cierta forma avanzada en comparación a la utilizada por la Teoría Clásica del Test. Era difícil examinar algunos de estos conceptos sin hacer una gran cantidad de cálculos para obtener información útil.

Por la razón antes expuesta no fue hasta los años sesenta del siglo pasado, con la aparición del libro de Rasch (1960) y, sobre todo, con los aportes de Bienbaun, Lord y Novick (1968) que se produce una rápida expansión en la utilización de la TRI, todo ello complementado con el acceso generalizado a los computadores, imprescindible para realizar con relativa facilidad los cálculos necesarios para el tratamiento de esta teoría.

La promesa central de la TRI fue solucionar las limitaciones de la TCT descritas en esta introducción, es decir:

- Obtener mediciones que no varíen en función del instrumento utilizado, que sean invariantes respecto de los tests empleados.
- Disponer de los instrumentos de medida cuyas propiedades no dependan de los objetos medidos, que sean invariantes respecto de los sujetos evaluados.

Adicionalmente la TRI proporciona todo un conjunto de avances técnicos que resultan de gran interés para la evaluación psicológica y la edumetría, tales como las funciones de información de los ítems y del test, los errores típicos de medida y una aplicación de la teoría de importancia capital para los sistemas de medición de la calidad de la educación: el establecimiento de bancos de ítems con parámetros estrictamente definidos.

Como se podrá apreciar a continuación, para lograr tales objetivos los supuestos de partida de la TRI son muy fuertes y restrictivos.

Supuestos de la TRI

Generalmente cuando se van a desarrollar acciones de medición psicológica y educacional, existe una variable fundamental de interés. Esa variable es conocida en la psicometría como "rasgo latente".

Un objetivo primario de la medición educacional y psicológica es la determinación de cuánto "rasgo latente" posee el individuo objeto de la medición.

Pero esos "rasgos latentes" en general no se pueden medir directamente como las dimensiones físicas, por ejemplo la altura y el peso.

Como en Educación la mayoría de las investigaciones han tratado a tales variables como habilidades (de lectura, aritmética, etc.), el término genérico de "habilidad" se usa dentro de la teoría de respuesta al ítem para referirse a estos rasgos latentes.

Normalmente para medir una "habilidad" se desarrolla un examen consistente en una cantidad determinada de ítems (preguntas). Cada uno de estos ítems mide alguna faceta de la habilidad de interés y la suma de los aciertos obtenidos por el examinando, llevada a cierta "escala" es el puntaje obtenido por él.

PRIMER SUPUESTO DE LA TRI:

Curva característica del item (CCI)

La TRI asume que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos, denominando a dicha función curva característica de los ítems (CCI). Ello significa que sujetos con distinta puntuación en dicha variable (en la prueba toda) tendrán posibilidades distintas de acertar determinado ítem.

Luego, una suposición razonable es que cada examinando que responda a un ítem de un examen, posee alguna cantidad de la habilidad medida por dicho examen. Por consiguiente, cada examinando tiene un valor numérico, una calificación, que lo coloca en la escala de habilidad. Esta habilidad es denotada por la letra griega θ.

Para cada nivel de habilidad, habrá cierta probabilidad de que un examinando con esa habilidad dará la respuesta correcta al ítem. Esta probabilidad es denotada por $P(\theta)$.

Para examinados de poca habilidad $P(\theta)$ será pequeña, mientras que para examinados con mucha habilidad $P(\theta)$ será grande.

En la figura 1 aparece la curva característica de un ítem. En el eje de abscisas se representan los valores de la variable que mide el ítem y en el de ordenadas aparece la probabilidad de acertar el ítem para los distintos valores de θ.

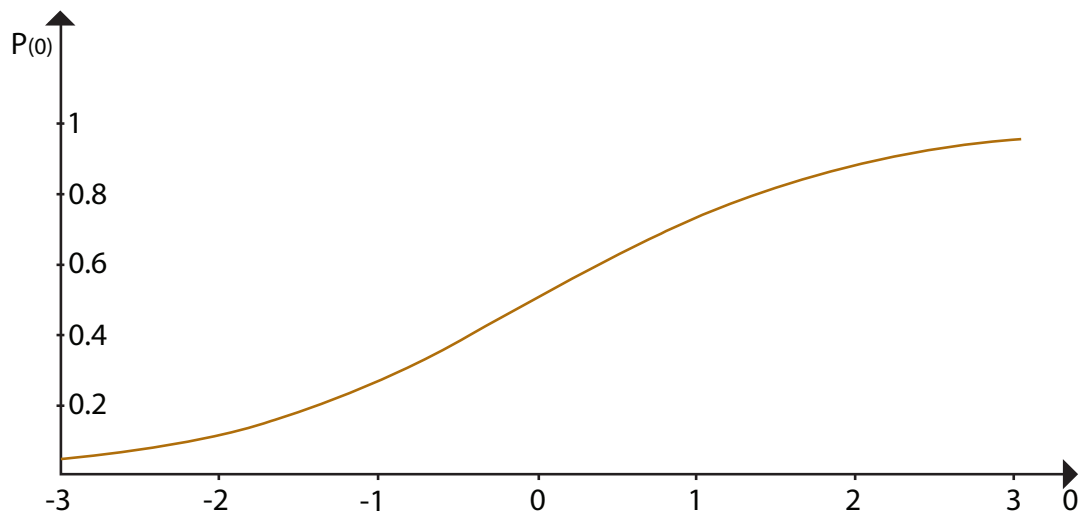


FIGURA 1: Curva Característica de un Ítem

Como se aprecia en el gráfico anterior esta curva en forma de S describe la relación entre la probabilidad de respuesta correcta a un ítem y la escala de habilidad. Esta última, si bien en la práctica muestra valores que van desde **-3** a **+3** , su margen teórico es desde el negativo infinito al positivo infinito.

La CCI, como su nombre lo indica, es eso, característica, típica, específica de cada ítem, caracteriza al ítem; por tanto, las CCI de los ítems que miden una determinada variable q no son iguales, si bien compartirán determinada forma general.

El margen restringido empleado en las figuras **(-3** a **+3)** es necesario solamente para ajustar las curvas a la pantalla de la computadora de forma razonable.

Parámetros de la CCI

La curva característica del ítem es la piedra angular de la teoría de respuesta al ítem; todas las otras estructuras de la teoría dependen de esta curva. Hay tres propiedades técnicas de la curva característica del ítem que la describen. Estas propiedades reciben el nombre de parámetros.

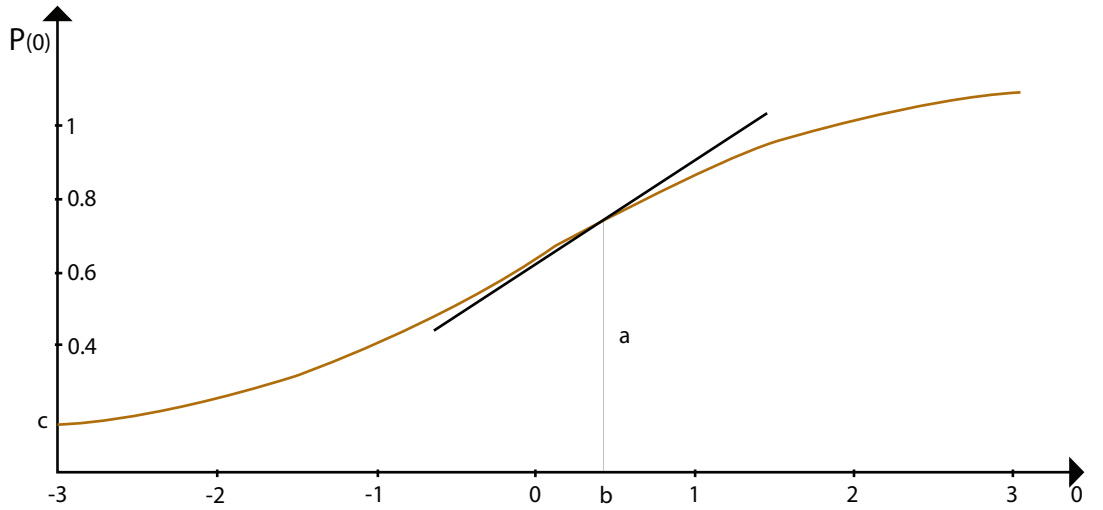


FIGURA 2. Parámetros de la CCI Parámetro a (índice de discriminación)

Parámetro a (índice de discriminación)

El valor de este parámetro es proporcional a la pendiente de la recta tangente a la CCI en el punto de máxima pendiente de esta.

Cuando mayor sea la pendiente, mayor será el índice de discriminación.

Esta propiedad técnica de la CCI describe cómo un ítem puede diferenciar entre los examinados que tienen habilidades inferiores a la localización del ítem y los que tienen habilidades superiores a la localización del ítem.

Mientras más pendiente tenga la curva, mejor se puede diferenciar el ítem mientras más llana sea la curva, menos puede diferenciar el ítem pues la probabilidad de repuesta correcta a niveles de habilidad bajos es casi la misma que en los niveles de habilidad altos.

A mayor discriminación la "**S**" es más pronunciada. A menor discriminación la "**S**" adopta una forma casi lineal, llana.

Parámetro b (Índice de dificultad)

En la teoría de respuesta al ítem, la dificultad del ítem describe dónde el ítem funciona en la escala de habilidad.

Por ejemplo, un ítem fácil funciona entre examinados de poca habilidad y un ítem difícil funciona entre los examinados de mucha habilidad. O sea este es un indicador de localización.

Nótese que en esta teoría la dificultad del ítem se mide en la misma escala que θ, de hecho es un valor de θ, aquel que corresponde a la máxima pendiente de la CCI, y en la práctica se puede obtener localizando el punto en el eje θ que corresponde a $P(\theta) = 0{,}5$, como puede verse en la figura 3.

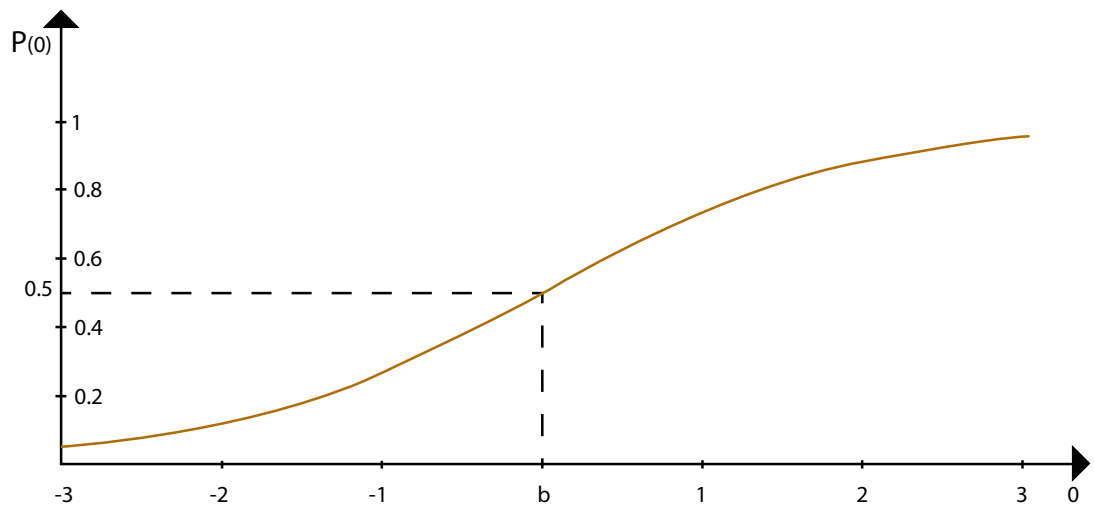


FIGURA 3. Ubicación del parámetro b

A continuación examinamos la idea de la dificultad del ítem como índice de localización.

En la figura 4, se presentan en el mismo gráfico tres curvas características del ítem. Todas tienen el mismo nivel de discriminación pero difieren con respecto a la dificultad. La curva de la izquierda representa un ítem fácil porque la probabilidad de respuesta correcta es alta para examinandos de poca habilidad y se acerca al **1** para examinandos de mucha habilidad. La curva del centro representa un ítem de dificultad media porque la probabilidad de respuesta correcta es baja en los niveles más bajo de habilidad, alrededor de **0,5** en el medio de la escala de habilidad y cerca de **1** en los niveles más alto de habilidad. La curva de la derecha representa un ítem difícil. La probabilidad de respuesta correcta es baja en la

mayor parte de la escala de habilidad y aumenta solamente cuando se alcanzan los niveles más altos de habilidad. Incluso en el nivel más alto de habilidad que se muestra en **(+3)**, la probabilidad de respuesta correcta es solamente **0,8** para el ítem más difícil.

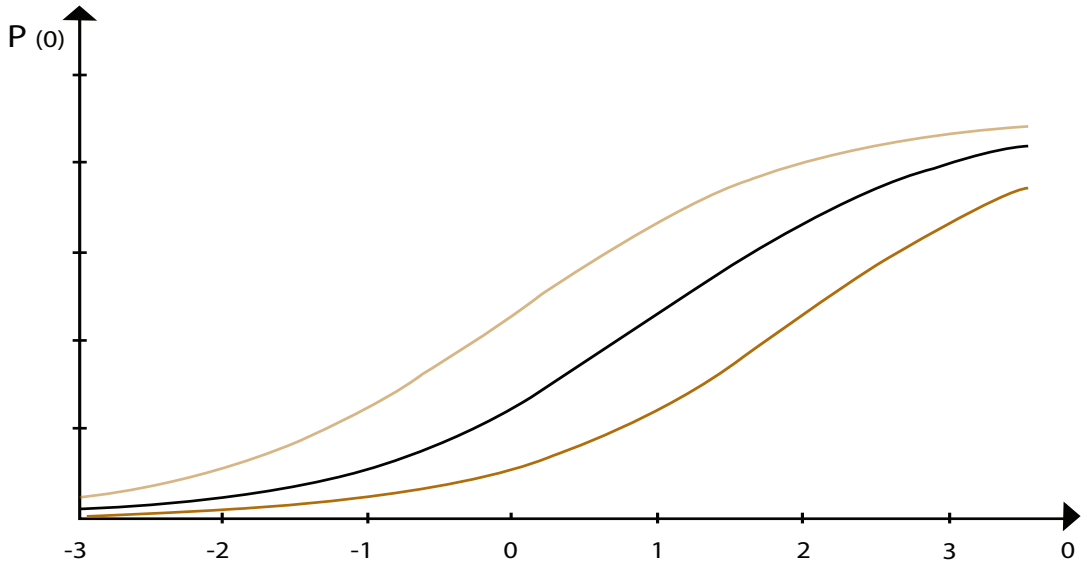


FIGURA 4. Tres CCI con la misma discriminación pero con diferentes niveles de dificultad.

El concepto de discriminación, se ilustra en la figura 5. Esta figura contiene tres curvas características de ítem que tienen el mismo nivel de dificultad pero defieren con respecto a la discriminación.

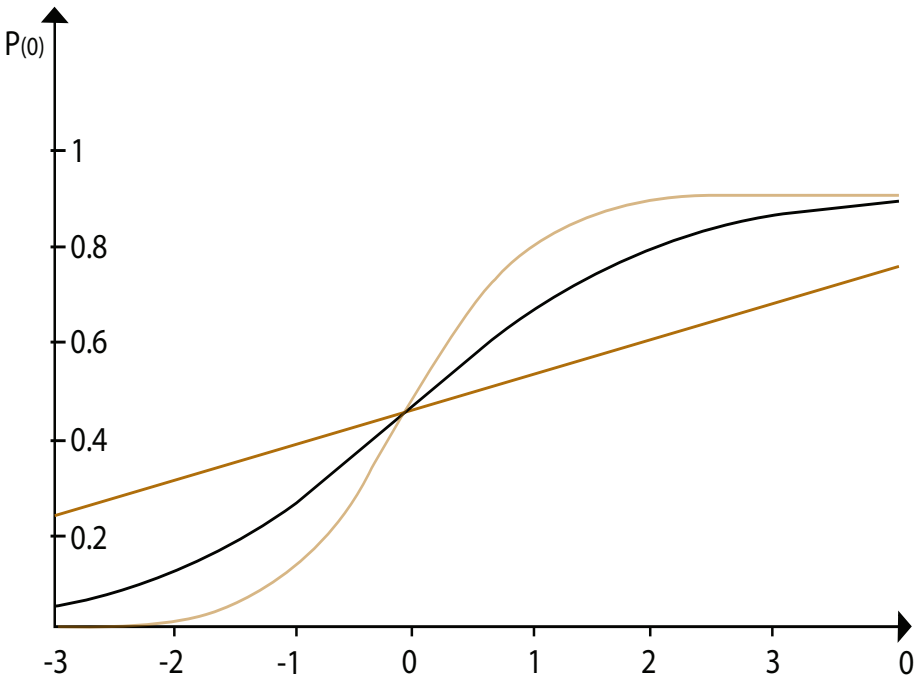


FIGURA 5. Tres CCI con la misma dificultad pero con diferentes niveles de discriminación.

La curva superior tiene un alto nivel de discriminación pues la curva tiene una gran pendiente en el medio en donde la probabilidad de respuesta correcta cambia muy rápidamente según aumenta la habilidad. Solamente a una pequeña distancia hacia la izquierda del medio de la curva, la probabilidad de respuesta correcta es mucho menor que **0,5**. La curva del medio representa un ítem con un nivel de discriminación moderado. El desnivel de esta curva es mucho menor que la anterior y la probabilidad de respuesta correcta cambia menos dramáticamente que la anterior según aumenta el nivel de habilidad.

Sin embargo, la probabilidad de respuesta correcta cambia menos dramáticamente que la anterior según aumenta el nivel de habilidad. No obstante, la probabilidad de respuesta correcta está cerca de cero para los examinandos de menor habilidad y cerca de 1 para los examinandos de mayor habilidad. La tercera curva representa a un ítem con poca discriminación. La curva tiene un desnivel pequeño y la probabilidad de respuesta correcta cambia lentamente por todo el margen de habilidades mostrado. Incluso en los niveles bajos de habilidad, la probabilidad de respuesta correcta es razonablemente grande y aumenta solo ligeramente cuando se alcanzan altos niveles de habilidad.

Parámetro c (pseudo adivinación)

El parámetro c representa la probabilidad de acertar el ítem al azar cuando "no se sabe nada", es decir, es el valor de $P(\theta)$ cuando $\theta = -\alpha$

En la práctica es el valor equivalente en el eje $P(\theta)$ interceptado por la CCI (Ver FIGURA 2).

La CCI queda definida cuando se especifican estos tres parámetros y se adopta una determinada función matemática para la curva. Según el tipo de función matemática adoptada y el valor de los parámetros tendremos diferentes modelos de CCI.

Tipos de modelos de CCI

En el apartado anterior se definieron las propiedades técnicas de la CCI en términos de descriptores verbales. Ciertamente los mismos son útiles para obtener una comprensión intuitiva de las curvas características del ítem, pero también debemos reconocer que carecen de la precisión y el rigor necesarios para una teoría.

Por tanto, en este epígrafe presentaremos tres modelos matemáticos para la curva característica del ítem.

Estos modelos proporcionan una ecuación matemática mediante la cual se relaciona la habilidad (θ) con la probabilidad de respuesta correcta $P(\theta)$. De esta manera dichos modelos y sus parámetros ofrecen un vehículo para comunicar información sobre las propiedades técnicas del ítem.

Hasta el momento la mayoría de las investigaciones que han abordado este tema, se han centrado en dos tipos de funciones matemáticas para la CCI: la función logística y la curva normal acumulada.

Dada la mayor "tratabilidad matemática" de la función logística, en la actualidad los tres modelos por antonomasia de la TRI son el logístico de un parámetro, de dos y de tres parámetros.

En los tres casos se asume que la respuesta a los ítems es dicotómica, es decir, o se acierta o se falla el ítem, independientemente del número de alternativas que tenga, o que sea de carácter abierto en el que los sujetos deben generar su propia respuesta, en cuyo caso ésta sólo se considerará correcta e incorrecta, sin grados intermedios. No obstante, en la literatura existen otros tipos de modelos para respuestas multicategoriales, pero no serán objeto de este curso.

La función logística

El objetivo de esta apartado es hacerles desarrollar a los cursistas un sentido sobre cómo se relacionan los valores numéricos de los parámetros del ítem para un modelo determinado con la forma de la curva característica del ítem.

Bajo la teoría de respuestas al ítem, el modelo matemático estándar para la curva característica del ítem es la forma acumulativa de la función logística. La misma define a una familia de curvas que tienen la forma general de las curvas características del ítem mostradas en el apartado anterior.

La función logística se derivó por primera vez en 1874 y ha sido ampliamente utilizada en las ciencias biológicas para hacer modelos del crecimiento de las plantas y animales desde el nacimiento hasta su madurez. Se utilizó por primera vez como modelo para la CCI a finales de los años cincuenta del siglo pasado y, por su simplicidad, se ha convertido en el modelo preferido.

Modelo logístico de un parámetro (modelo de Rasch)

El modelo logístico de un parámetro fue formulado originalmente por Rasch (1960), recibiendo notable atención desde entonces especialmente en la Universidad de Chicago por Wright y Stone.

Es, sin dudas, el modelo más popular de la TRI debido esencialmente a la sencillez emanada de su lógica: la respuesta a un ítem sólo depende de la competencia del sujeto (θ) y de la dificultad del ítem **(b)**. En este modelo la CCI viene dada por la función

$$Y = \frac{e^{x}}{(1+e^{x})}$$

logística, y el único parámetro de los ítems a tener en cuenta es **b** (índice de dificultad). La función logística es una curva cuya fórmula general viene dada por:

donde: **e:** base de los logaritmos neperianos, o sea, **e=2,7182... =2,72**

Ejemplo:

X	Y
-α	0,0
-3	0,047
-2	0,119
-1	0,269
0	0,500
1	0,731
2	0,881
3	0,953
+α	1,0

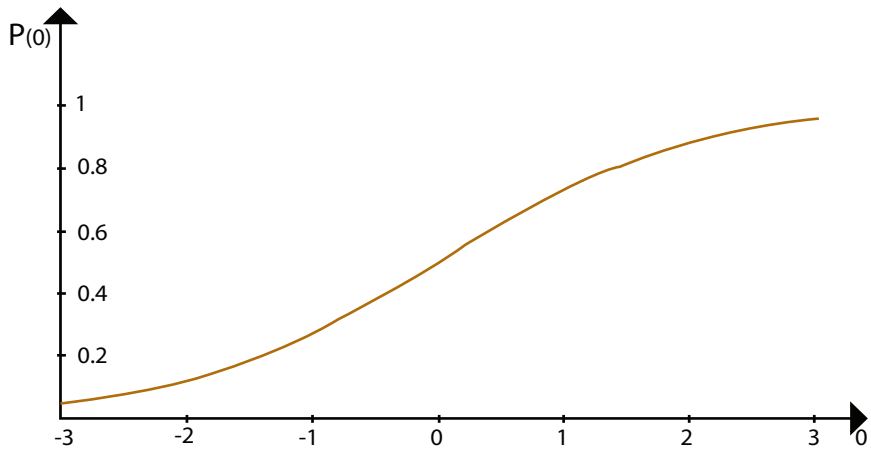


FIGURA 6. Gráfica del modelo logístico de un parámetro

Mediante el uso de una constante adicional (**D = 1,7**) en la función logística sus valores se aproximan notablemente a los de la curva normal acumulada, por lo que es frecuente encontrarla como sigue:

$$P_i(\theta) = \frac{e^{Dx}}{(1+e^{Dx})}$$

que, adaptada a la terminología de la TRI para el caso particular de un parámetro, en el modelo de Rasch la CCI adquiere la expresión siguiente:

$$P_i(\theta) = \frac{e^{D(\theta-b_i)}}{1+e^{D(\theta-b_i)}}$$

donde: **P(θ) :** Probabilidad de acertar el ítem i para un nivel de habilidad θ.

θ: Valores de habilidad en la variable medida.

$\mathbf{b_i}$: Índice de dificultad del ítem **i**.

e: Base de los logaritmos naturales (**2.72**)

D: Constante (**1,7**)

Este modelo significa que:

- Conocido el índice de dificultad del ítem (**b**).
- Y la competencia de los sujetos θ
- El modelo predice la probabilidad P(θ) de que acierten el ítem.

Nota: En adelante prescindiremos del subíndice **i**.

Ejemplo: 1 ¿Cuál es la probabilidad de que los sujetos con $\theta=2$ acierten un ítem cuyo índice de dificultad es **b = 1,5**?

$$P_i(\theta)= \frac{2{,}72^{1{,}7(2-1{,}5)}}{1+2{,}72^{1{,}7(2-1{,}5)}}$$

$$P_i(\theta)= \frac{2{,}72^{0{,}85}}{1+2{,}72^{0{,}85}}$$

$$P_i(\theta)= \frac{2{,}34090381}{1+2{,}34090381}$$

$$P_i(\theta)= 0{,}7$$

La fórmula dada para el modelo de Rasch suele expresarse de una manera equivalente, resultado de dividir al numerador y al denominador de esta por $e^{D(\theta-b)}$; en cuyo caso quedaría expresado como:

$$P_i(\theta)= \frac{1}{1+e^{-D(\theta-bi)}}$$

Modelo logístico de dos parámetros

El modelo logístico de dos parámetros fue originariamente desarrollado por Bienboum (1957, 1958, 1968). Asume que la CCI viene dada por la función logística y contempla dos parámetros de los ítems, el índice de discriminación (**a**) y el índice de dificultad (**b**). Su fórmula viene dada por la expresión:

$$P_i(\theta)= \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

Donde: $P(\theta)$: : Probabilidad de acertar el ítem **i** para un valor **θ**.

θ : Valor de la variable medida.

$\mathbf{b_i}$: Índice de dificultad del ítem **i**

$\mathbf{a_i}$: Índice de discriminación del ítem **i**.

e: Base de los logaritmos referíamos (**2,72**)

D: Constante (**1,7**)

Ejemplo:

El índice de discriminación de un ítem es 2, y su índice de dificultad 1,5: ¿Qué probabilidad tienen de acertar ese ítem los sujetos cuyo nivel de habilidad en la variable medida sea 2,5?

Datos: **a=2**; **b=1,5**; **θ=2,5**; **D=1,7**

$$P(\theta)= \frac{(2,72)^{(1,7)(2)(2,5-1,5)}}{1+ (2,72)^{(1,7)(2)(2,5-1,5)}} =0,967$$

La probabilidad de superar el ítem es muy elevada **(0,967)**, como era de esperar, pues a medida que sea mayor que b, para un determinado valor de a, **P(θ)** aumenta según el modelo logístico, lo cual es razonable, pues a mayor habilidad de los sujetos, mayor probabilidad de superar un ítem dado.

Modelo logístico de tres parámetros

Este modelo asume que la CCI viene dada por la función logística y añade a los dos parámetros **a** y **b** ya citados un tercero **c** relativo a la probabilidad de acertar el ítem al azar, cuando no se conoce la respuesta. Mas técnicamente, **c**, es el valor de $\mathbf{P_i(\theta)}$ para un valor de $\theta = -\alpha$

En la práctica **c**, es el intercepto de la curva con el eje **P(θ).**

El modelo puede expresar como sigue:

$$P_i(\theta)=c_i + (1-c_i) \frac{e^{Da_i(\theta-b_i)}}{1+ e^{Da_i(\theta-b_i)}}$$

Ejemplo:

La probabilidad de acertar cierto ítem al azar es **0,25**, su índice de dificultad es **0,5** y su índice de discriminación es **1,25**. ¿Cuál es la probabilidad de acertar ese ítem para sujetos con **q =1**?

Datos: **θ=1**; c_i**=0,25**; **D=1,7**; a_i**=1,25**; b_i**=1,5** ;;

$$P_i(1)=0,25+(1-0,25) \frac{(2,72)^{(1,7)(1,25)(1-0,5)}}{1+ (2,72)^{(1,7)(1,25)(1-0,5)}} 0,805$$

$P_i(1)=0,805$

Nota: Como seguramente ya el lector ha comprendido, el modelo logístico de tres parámetros es el más general: si se hace **C=0**, se obtiene el de dos parámetros y si además **a** se asume constante para todos los ítems se obtiene el de un parámetro.

SEGUNDO SUPUESTO: Unidimensionalidad.

Como ya hemos explicado anteriormente la CCI establece una relación funcional entre la probabilidad de acertar un ítem y los valores de θ. Por tanto, si el modelo es correcto, la probabilidad de acertar un ítem únicamente dependerá de un factor, de θ.

En otras palabras, la TRI asume en su formulación que los ítems destinados a medir la variable θ constituyen una sola dimensión, son unidimensionales.

Sobre cómo comprobar que un conjunto de ítems constituye una sola dimensión existen diversas opiniones entre los investigadores, habiéndose propuesto hasta la fecha números índices al respecto. No obstante, el análisis factorial sigue siendo el método más utilizado.

Dado que empíricamente raras veces se encuentra una unidimensionalidad perfecta, o sea, que un solo factor dé cuenta del **100%** de la varianza, la unidimensionalidad en la práctica se verifica a partir de cuánta más varianza explique el primer factor.

Un problema clásico de difícil solución que surge al someter a un análisis factorial ítems dicotómicos, como son la mayoría de los utilizados en los tests que aplican todos los consorcios internacionales de evaluación, es lo que se ha dado en llamar "factores de dificultad", refiriéndose a que los factores obtenidos dependen en cierta medida de la dificultad de los ítems.

TERCER SUPUESTO: Independencia local

Del cumplimiento de la unidimensionalidad se deriva matemáticamente la existencia de independencia local. Esto significa que para un sujeto con un determinado valor en la variable unidimensional su respuesta a un ítem no está influida por su respuesta en los otros ítems.

La independencia local puede expresarse de otra manera, diciendo que la probabilidad de que un sujeto acierte n ítems es igual al producto de las probabilidades de acertar cada uno de ellos.

Ejemplo:

Si un test consta de tres ítems y la probabilidad de acierto de un sujeto en el primer ítem es $\mathbf{P(A_1)=0,40}$, de que acierte el segundo $\mathbf{P(A_2)=0,50}$ y el tercero $\mathbf{P(A_3)=0,80}$, lo que establece el principio de independencia local es que la probabilidad de que este sujeto acierte los tres ítems es:

$$\mathbf{P(A1, A2, A3) = (0,40)(0,50)(0,80)=0,016}$$

Analógicamente puede hablarse de independencia local de los sujetos en el sentido de que el rendimiento de un sujeto es independiente del rendimiento de los otros.

Comprobación del ajuste a los datos de los modelos.

¿Cómo proceder en la práctica para elegir uno de los modelos, estimar los parámetros de los ítems, la habilidad de cada sujeto y comprobar que el modelo se ajusta a los datos?

A continuación se describe el algoritmo que debe seguirse para dar respuesta a la pregunta anterior, explicando brevemente en qué consiste cada uno de los pasos o etapas por las que se debe transitar.

1. Definición rigurosa de la variable que se pretende evaluar.

Este primer paso no es específico de la TRI, atañe a cualquier medición psicológica o educacional rigurosa.

Si no se define con rigor aquello que se desea evaluar, mal se podrá medir.

Definir con rigor no se refiere únicamente a dejar claros los "deseos", lo que se pretende medir, hay que delimitar el marco teórico, las posibles conexiones con otras variables y teorías, antecedentes, etc.

En tal sentido la medición podrá hacerse:

- Con arreglo al estado del arte a nivel mundial de la dimensión disciplinar que se pretende evaluar.
- Para evaluar el estado de la dimensión disciplinar en su aplicación estrictamente curricular o un subconjunto de esta (precisando la oportunidad real de aprendizaje que han tenido los alumnos).
- Una mezcla de las variantes anteriores (Ejemplo: dimensión disciplinar curricular con énfasis en las habilidades para la vida).

En cualquier caso, resulta necesario operacionalizar el marco teórico precisamente definido a través de una "tabla de especificaciones". L a prueba a elaborar deberá entonces satisfacer la referida tabla.

2. Elaboración de los ítems destinados a medir la variable.

Elaborar buenos ítems es como escribir poesías, si existieran reglas de aplicación automática todos seríamos excelentes poetas, pero desafortunadamente no es así. No obstante, a continuación mostramos algunas reglas para la elaboración de ítems de selección múltiple con única respuesta correcta, cuya observancia puede ayudar sin dudas a elaborar ítems con una adecuada calidad:

Algunas consideraciones sobre la elaboración de preguntas para las pruebas pedagógicas de lápiz y papel.

Actualmente la mayoría de los consorcios internacionales que hacen evaluación educativa comparada tales como TIMSS, PISA, LLECE, SALMEQ, etc., utilizan mayoritariamente en sus pruebas preguntas o ítems con formato de selección múltiple con una única respuesta correcta, especialmente por la facilidad de su aplicación y calificación. Sin embargo, es errado creer que resulta fácil también su elaboración, todo lo contrario.

Elaborar buenos ítems es una labor que exige mucha práctica, además de un conocimiento profundo del objeto de evaluación y de la población que se va a evaluar.

Para que cada ítem aporte su máximo potencial al propósito de la evaluación resulta necesario que sus elaboradores dominen el marco conceptual del instrumento, así como las especificaciones técnicas del mismo.

El proceso de diseño y aplicación de pruebas integra diferentes, a saber:

- Elaboración de su fundamentación conceptual.
- Construcción de la tabla de especificaciones de la prueba.
- Desarrollo de los ítems y de la prueba en su conjunto.
- Pilotaje.
- Aplicación.
- Análisis de ítems y de la prueba en su conjunto.
- Escala de calificación.
- Producción y divulgación de resultados.

Obviamente, luego de construir el marco teórico de la prueba, debe procederse a la elaboración de ítems, en cuyo proceso deben tenerse en cuenta las siguientes reglas. [2]

Reglas generales

- Verificar que el ítem corresponda con los propósitos de la evaluación, la estructura de la prueba y con las dimensiones disciplinares. Todas las preguntas de una prueba deben ser independientes entre sí. La información de un ítem no debe servir de pauta para contestar otro, ni la respuesta a un ítem debe depender de haber encontrado primero la de otra anterior.
- Evitar los ítems que pueden contestarse por sentido común y aquellos cuya respuesta dependa únicamente de recordar un término, un símbolo, un dato o la fecha en que ocurrió un evento.
- Evitar expresiones rebuscadas que puedan confundir. Se recomienda emplear un lenguaje directo, sencillo y comprensible.
- Los ítems no deben tener juicios de valor explícito o implícito.

2 Rocha, Martha y otros. Seminario regional Evaluación de la Educación. Taller de elaboración de ítems, ICFES, agosto de 2006, p. 10.

Reglas sobre los enunciados

- Los enunciados deben ser afirmativos, en caso de ser necesaria la negación, se debe resaltar para llamar la atención hacia la formulación negativa. La doble negación afecta la comprensión ("No es cierto que no procedan los recursos").
- Evitar enunciados demasiado extensos y poco atractivos ya que desmotivan la lectura, disminuyen el tiempo de respuesta y fatigan.
- Garantizar la coherencia interna del enunciado y de este con las opciones de respuesta.

Reglas sobre las opciones

- Las opciones de respuesta deben pertenecer al mismo campo semántico.
- Las opciones de un ítem no deben dar indicaciones sobre la clave por ofrecer un cierto contraste evidente de:
- longitud precisión / imprecisión
- uso común / técnico
- generalización / particularización
- No se deben repetir expresiones en las opciones de respuesta, si éstas se pueden incluir en el enunciado del ítem. Repetir la misma palabra del enunciado en cualquiera de las opciones lleva a que sea elegida como respuesta, sin serlo necesariamente.
- Debe evitarse en las opciones las expresiones "todas o ninguna de las anteriores", en su lugar es necesario construir alternativas de respuestas plausibles para las personas que no tengan el dominio conceptual que exige el ítem.
- Realizar una revisión gramatical y ortográfica de cada uno de los ítems.

Fases del diseño de la prueba

Fundamentación conceptual

Esta es la fase en la que se aborda conceptualmente el objeto de la evaluación. En ella básicamente se da respuesta a las interrogantes siguientes:

- ¿Para qué la evaluación? (Su propósito)
- ¿Qué se evalúa? (Su objeto)
- ¿A quién se evalúa? (Características de la población objetivo)
- ¿A quién le será útil la evaluación y de qué manera?(Usuarios de la evaluación y precisión de los beneficios de la misma)

Especificaciones de la prueba

Constituyen una descripción lo más detallada posible de las características del instrumento. Suele utilizarse para hacer tal descripción una tabla de doble entrada: en la primera columna es común colocar una desagregación del objeto de evaluación en dimensiones (tópicos disciplinares) y en el resto de las columnas se suele escribir la delimitación de las especificidades técnicas del instrumento, tales como longitud (estructura de la prueba) y formato (especificidades psicométricas).

Al hacer el análisis de una prueba que satisfaga la "tabla de especificaciones" antes referida es posible describir, diagnosticar, el rendimiento de los alumnos que la realicen desde el punto de vista conceptual, procedimental y actitudinal. Puede incluso construirse un índice con cada una de sus dimensiones (conceptual, procedimental y actitudinal) y precisar en cuál de ellas los alumnos tienen mayores deficiencias.

Particular importancia tiene el poder constatar el grado de asimilación que han alcanzado los alumnos en la dimensión actitudinal, la cual sin dudas es un componente esencial de sus orientaciones valorativas.

La descripción de las tareas de evaluación a partir de las cuales será posible materializar (en ítems) el propósito de la evaluación constituye la operacionalización del objeto de evaluación.

Como para de las especificaciones psicométricas se debe clarificar el número de ítems que tendrá el instrumento en su totalidad.

Desarrollo de la prueba

Es esencial destacar que el propósito de esta fase es producir un instrumento de evaluación y no un agregado de ítems.

Después de elaborar los ítems, se ensambla el instrumento de acuerdo con las especificaciones psicométricas y se somete a una revisión final, generalmente por parte de expertos en evaluación y en el objeto de evaluación.

De manera que una prueba es un conjunto intencionalmente articulado de ítems a través de cuya aplicación se infiere el desempeño de quienes son evaluados en relación con el objeto de la evaluación.

Aplicación piloto

En esta fase el instrumento es aplicado a una muestra de la población objetivo para estimar indicadores estadísticos que permitan corroborar la calidad técnica del instrumento y el grado de pertinencia de los ítems para la población.

Las condiciones de la aplicación piloto deben guardar la mayor similitud posible con las condiciones que tendrá la aplicación definitiva.

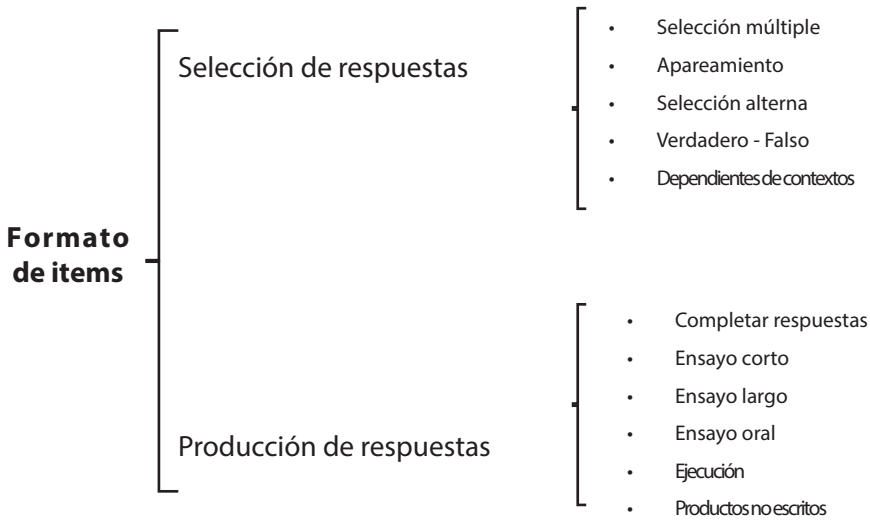
Cuando no existen las condiciones logísticas necesarias para garantizar una adecuada aplicación piloto de la prueba, se puede utilizar como alternativa un "*juicio de expertos*", o sea un procedimiento a través del cual un equipo de personas expertas en el objeto de evaluación y en procesos de evaluación, los que califican los ítems de un instrumento a la luz de las consideraciones planteadas en la fundamentación conceptual de la prueba y su coherencia con los objetivos de la evaluación, sus especificaciones psicométricas y las características de la población objetivo.

Contenido de los ítems

Para hacer ítems de buena calidad es necesario conocer las características de los distintos formatos de ítems que han de utilizarse en la evaluación, aplicar las reglas para su correcta elaboración y evitar los factores que pueden afectar su validez.

Posibles formatos de los ítems

A continuación mostramos una clasificación de formatos de ítems tomando como criterio o base de la misma lo que el evaluado hace cuando se enfrenta a un ítem, o sea seleccionar o producir una respuesta.



En la elaboración de un instrumento es posible incluir variedad de formatos. Ahora bien, cada tipo de formato tiene un conjunto de requisitos particulares de elaboración y tiene distinto potencial evaluativo.

Generalmente se procura incorporar un número significativo de ítems de un mismo formato a los efectos de que la lectura de las instrucciones no relegue a un segundo plano el propósito evaluativo fundamental.

En este libro solo mostraremos la técnica de elaboración de ítems del formato "selección múltiple con una única respuesta correcta" y en la ejemplificación pondremos el énfasis en aquellos ítems cuyo propósito es evaluar los contenidos actitudinales.

El ítem de selección múltiple con única respuesta correcta

Las partes que componen un ítem de este tipo son:

- El contexto
- El enunciado
- Las opciones de respuestas

El contexto

Es la información que sitúa conceptualmente al evaluado pues provee los elementos necesarios y suficientes para focalizar la tarea de evaluación. Este puede ser un texto, una gráfica, un dibujo, una tabla o cualquier otra forma de presentación de la información a partir de cual se deriva el enunciado.

El enunciado

Es el planteamiento propiamente dicho, de la problemática que se espera sea resuelta por el evaluado. [3]

Comúnmente los enunciados de ítems de selección múltiple como única respuesta correcta se plantean en forma de pregunta o como una proposición. En el primer caso las opciones se redactan como respuestas a la pregunta; en el segundo caso, en enunciado constituye la primera parte de una proposición y cada una de las opciones debe completar coherentemente el enunciado.

Es conveniente tener en cuenta que para los niños la redacción en forma de preguntas resulta mas clara.

Las opciones de respuesta

Las opciones son posibles inverosímiles respuestas a la problemática planteada en el enunciado.

Reglas para la elaboración de ítems de selección múltiple con única respuesta correcta.

Las reglas de elaboración de ítems que a continuación les presentamos, tienen la pretensión de orientar al evaluador para que los ítems que elabore evalúen realmente el objeto de evaluación que se pretende; en tal sentido, la aplicación de tales reglas contribuye a consolidad la validez de la interpretación de los resultados[4].

3 AERA, APA, NCME (1999) standard for educational and psychological testing. Washintong: AERA

4 Rocha, Martha y otros. Seminario regional Evaluación de la Educación. Taller de elaboración de ítems, ICFES, agosto de 2006, pág. 30-31.

Reglas sobre el contenido de los ítems

- Evite elaborar ítems que confunden al evaluado. Diferentes estudios han establecidos cuáles son algunas de las situaciones que llevan a percibir los ítems como confusos entre éstas están:

a) Contenido trivial.
b) Presencia de información irrelevante.
c) Presentación ambigua de las opciones de respuesta.
d) Discriminación muy fina – difícil de percibir entre las opciones de respuesta.
e) Presentación de información en modo distinto a como ha sido aprendida por la población evaluada, dentro de su proceso educativo.

- Cada ítem debe corresponder a una tarea de evaluación definida en al estructura de prueba.
- Evite evaluar el mismo aspecto específico con varios ítems. Aproveche cada ítem para hacer cada vez más completa la evaluación.
- Plantee una sola problemática en cada ítem.
- Evite ítems que incluyan posiciones ideológicas o prejuicios; que tenga en cuenta que las proposiciones prejuiciosas pueden resultar en una ofensa para cualquiera de los evaluados. Se exceptúa esta recomendación si justamente dichas posiciones son el objeto de evaluación; entonces será obligatorio incluirlas.
- El vocabulario utilizado debe ser adecuado para la población objetivo.
- Cada ítem debe ser independiente y no proveer información para responder a otros.
- No utilice ítems que aparezcan en libros, revistas u otros documentos, como base para sus ítems. Elabore ítems originales.
- Evite ítems en los cuales se indague la opinión (parecer no argumentado) del evaluado (a menos que el instrumento justamente pretenda servir para un sondeo de opinión).
- Evite plantear ítems cuya respuesta válida se determine según la opinión de quien la elabora.
- Balancee la complejidad de los ítems para que el instrumento cubra los niveles de habilidad de la población objetivo, es decir, la prueba debe incluir ítems de dificultad alta, media y baja.

Reglas sobre construcción del enunciado

- Si plantea el enunciado en forma de proposición incompleta asegúrese de usar conjugaciones verbales, género y número adecuados para las opciones de respuesta que planteará. Si lo escribe en forma de pregunta asegúrese de usar adecuadamente signos de interrogación y la estructura gramatical de una pregunta.
- Presente en el enunciado la tarea de evaluación.
- Escriba con claridad.

- Evite texto excesivo.
- Redacte el enunciado en forma positiva; es decir, evite negaciones.

Reglas sobre construcción de opciones de respuesta

- Asegure la concordancia gramatical entre la proposición del enunciado y cada opción.
- Organice las opciones en un orden lógico (alfabético, longitud, etc.) o numérico.
- Mantenga la independencia entre las opciones. Éstas no deben solaparse o intersectarse y no deben ser sinónimas.
- Refiérase en todas las opciones al problema planteado en el enunciado. Evite opciones fácilmente descartables.
- Elabore opciones de respuesta de longitud similar.
- Evite colocar como opción:
  - Todos los anteriores
  - Ninguno de los anteriores
  - A y B son correctas (o cualquier combinación de opciones)
  - No sé
- Redacte las opciones en forma positiva, es decir evite negaciones. Si debe colocar una negación, resáltela (use negrilla o mayúsculas sostenidas).
- No repita en las opciones frases contenidas en el enunciado.
- Elabore ítems con 4 opciones de respuesta. Elaborar opciones plausibles es dispendioso; seguramente ganará calidad en las que redacte si no son demasiadas. Hay referencia de distintos estudios que analizaron la cantidad de opciones útiles para los propósitos de evaluación; si bien no existe consenso alrededor de un único número de opciones, se encuentra a menudo conveniente, en cuanto a facilidad de redacción y capacidad de discriminación, trabajar con 4 opciones, para poblaciones de infantes puede ser conveniente usar 3 opciones.
- Evite en las opciones el uso de adverbios como:
  - Siempre
  - Nunca
  - Totalmente
  - Absolutamente
  - Completamente
- La posición de la opción válida debe balancearse entre todos los ítems del instrumento. Es recomendable que aparezca proporcionalmente en cada posición posible.
- Evite que la opción válida pueda ser identificada fácilmente por contraste con las demás opciones, por alguna de las siguientes situaciones:
  - Tener mayor longitud
  - Ser la proposición de mayor precisión o imprecisión

- Estar redactada en un tipo de lenguaje diferente (técnico o común)
- Tener el mayor nivel de generalización o de particularidad
- Tener las mismas palabras que el enunciado
- Referirse a una problemática o tema diferente

• Justifique adecuadamente cada una de las opciones para garantizar que sólo hay una válida que las demás son plausibles para quienes no dominan completamente la tarea de evaluación.

Sobre la validez

Factores que afectan la validez

Cuando hablamos actualmente de validez no nos referimos al instrumento, sino a las inferencias e interpretaciones realizadas a partir de los resultados obtenidos en un proceso de evaluación donde se ha utilizado el instrumento en cuestión.

Entendemos entonces por validez "el juicio evaluativo del grado en el cual la evidencia empírica sustentan la pertinencia y conveniencia de las inferencias acerca de los resultados en un instrumento de medición así como de las acciones que se realizan a partir de dichos resultados". [5]

La elaboración de los ítems puede verse afectada en cuando a la validez por los siguientes factores:

- La tarea planteada por el ítem no es relevante para la evaluación del objeto planteado en el marco de fundamentación.
- En el ítem se incluye información que facilita o dificulta la tarea de evaluación planteada, más allá de su propósito.
- No se garantiza la confidencialidad del instrumento antes de su aplicación.

No existe una fórmula única y universal para mejorar la calidad de un ítem, pero sin lugar a dudas el estricto cumplimiento de las reglas para su elaboración y el tomar distancia de los factores que pueden afectar su validez, ayudan de manera decisiva a conseguir que tengan una adecuada calidad.

3. Aplicación de los ítems a una muestra amplia de sujetos perteneciente a la población en lo que se utilizará el futuro test y cálculo de los índices clásicos de los ítems.

Los ítems elaborados se aplican a una muestra lo más amplia posible de sujetos pertenecientes a la población en la que se va a utilizar (pilotaje), y se calculan para cada ítem los índices de la Teoría Clásica del Test, lo cual permitirá hacer una primera decantación de algunos que resultan claramente inadecuados. Programas con el ITEMAN para computadores personales facilitan esa labor.

5 Messick, S (1989). Validity. In R. L. Linn (Ed.). Educational measurement (3rd ed. Págs 13 – 103. New York: Mac millan)

4. Comprobación de la unidimensionalidad de los ítems.

El análisis factorial sigue siendo la técnica más apropiada, pero no hay un criterio claro a partir del cual se puede afirmar la unidimensionalidad. No obstante, el porcentaje de varianza explicada por el primer factor es un índice sencillo y claro de la relevancia del factor y, por ende, del grado de unidimensionalidad.

El proceso de unidimensionalidad de los ítems suele realizarse en varios pasos:

- Un primer análisis factorial que descarta aquellos ítems que conforman factores periféricos.
- Se hace lo mismo en posteriores análisis hasta lograr un análisis en el que un factor explica la mayor parte, idealmente toda, de la varianza de los ítems.

5. Elegir uno de los modelos de TRI

Una vez probado que los ítems conforman un test unidimensional, el siguiente problema es qué modelos de TRI es más razonable utilizar.

Cualquier elección a priori es lícita para el investigador, pero será el ajuste del modelo a los datos lo que decida lo correcto o incorrecto de la elección.

Ahora bien, ciertas características de los ítems pueden proporcionar algunas claves que mejoren la mera elección al azar o capricho, entre estas:

- es poco razonable intentar ajustar un modelo de un parámetro (Rasch) si se sospechan índices de discriminación no iguales, lo cual puede evaluarse tentativamente escrutando dichos índices en la Teoría Clásica del Test (TCT), o si es alta la probabilidad de acertarlos al azar (el modelo de un parámetro asume un índice de discriminación constante (**a=K**) y (**c=0**).
- Si **c=0** el modelo de dos parámetros es poco plausible a priori.
- Técnicamente, el modelo de tres parámetros debería ser preferible a los de uno y dos, pues estos constituyen casos particulares de aquel. Sin embargo, el de un parámetro es de cálculo e interpretación sencillos, por lo que en la práctica es el preferido de la mayoría de los usuarios. Incluso es atractivo desde el punto de vista teórico por su parsimonia, al postular que la respuesta de un sujeto a un ítem solo depende de su habilidad en la variable medidas por el ítem (**q**) y de la dificultad del ítem (**b**).

Por otra parte, la estimación del parámetro c en el modelo de tres parámetros no es un asunto totalmente resuelto aún.

En todo caso, preferencia aparte, no debemos olvidar que los jueces han de ser los datos y que se debe elegir aquel modelo que mejor de cuenta de ellos.

En caso de ajustes similares debe escogerse el más sencillo, como indican los cánones de la parsimonia científica y el sentido común (lo bueno, si es sencillo, es dos veces bueno).

6. Estimación de los parámetros de los ítems y de la habilidad de cada sujeto en la variable medida (θ).

Seleccionado uno de los modelos, el paso siguiente será estimar los parámetros de cada ítem y el valor de la variable medida (θ) para cada sujeto a partir de los datos obtenidos al aplicar los ítems a una muestra amplia de sujetos (pilotaje),

La estimación se va haciendo por aproximaciones sucesivas (iteraciones) y su cálculo es muy laborioso, por lo que son necesarios los computadores. El proceso de iteraciones se detiene cuando los valores estimados de los parámetros convergen, o sea, cuando tras una iteración no se producen cambios significativos en los valores estimados.

Actualmente se dispone de varios programas de ordenador para estos fines, entre ellos.

- BICAL (Wright, 1979) para modelos logísticos de un parámetro.
- BILOG (Michuy y Bock, 1984) para modelos logísticos de uno, dos y tres parámetros)

Estos programas ofrecen como salida fundamental los valores estimados de los parámetros de cada ítem y el valor de θ de cada sujeto.

7. Comprobar que el modelo se ajusta a los datos.

Una vez estimados los parámetros del modelo debemos comprobar hasta qué punto los resultados pronosticados con esos valores coinciden con los obtenidos de hecho, o sea, hay que comprobar el ajuste del modelo a los datos.

Tal ajuste se produce cuando los valores de **P(θ)** pronosticados por el modelo no defieren estadísticamente de los obtenidos empíricamente, es decir, de la proporción de sujetos que realmente acierten el ítem.

Existen varios procedimientos estadísticos para la comprobación del ajuste, si bien ninguno de ellos es totalmente satisfactorio, siendo esto precisamente un punto débil en el estado actual de desarrollo de la TRI.

Tres métodos antes referidos son:

- El uso de $\mathbf{X^2}$ (chi-cuadrado).
- El análisis de los residuos.
- La comparación de las distribuciones de las puntuaciones.

Curva Característica del Test (CTC)

De la misma manera que existe en la TRI el concepto de CCI, el que constituye una pieza central de dicha teoría, puede hablarse en la misma de curva característica del test (**CCT**), concepto que tiene también gran importancia, sobre todo porque constituye un puente entre algunos aspectos de la Teoría Clásica del TEST (TCT) y la TRI, como ayuda para interpretar los resultados, o en la equiparación de las puntuaciones de los sujetos (equating).

La curva característica del test es la suma de las curvas características de los ítems que componen el test, o sea, si a cada nivel de θ se suman los valores de $P(\theta)$ de cada ítem para ese nivel, se obtiene la **CCT**, lo que puede expresarse matemáticamente como sigue:

$$\mathbf{CCT} = \sum_{i=1}^{n} \mathbf{P_i}(\theta)$$

siendo "**n**" el número de ítems de la prueba o test.

Resulta necesario hacer notar que las sumas han de realizarse para cada nivel de **θ** y dado que **θ** es una variable continua, habría que utilizar el cálculo infinitesimal, si bien en la práctica es habitual dividir **θ** en cortos intervalos sumando la **P(θ)** de los ítems para cada intervalo.

Ejemplo:

Dado un test formado por 4 ítems cuyos parámetros en un modelo logístico de dos parámetros estimados con determinado programa de computación resultaron: $\mathbf{a_1=1}$; $\mathbf{a_2=1,5}$; $\mathbf{a_3=2}$; $\mathbf{a_4=2,5}$; $\mathbf{b_1=0,75}$; $\mathbf{b_2=1}$; $\mathbf{b_3=2}$ y $\mathbf{b_4=3}$. Hallar la curva característica del test (CCT). Hacer la suma de las **P(q)** para los valores de **θ**: **-3**, **-2**, **-1**, **0**, **1**, **2**, **3**.

Para dar respuesta al ejercicio anterior, sólo habría que sustituir los valores dados de **a**, **b** y **θ** en el modelo logístico de dos parámetros, obtener los valores de **P(θ)** para los 4 ítems, y sumar sus resultados para obtener la CCT.

Hagamos como ejemplo los cálculos para $\mathbf{\theta=1}$; $\mathbf{a_1=1}$ y $\mathbf{b_1=0,75}$:

$$\mathbf{P_i(\theta)} = \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

$$\mathbf{P_i(1)} = \frac{e^{(1,7)(1)(1-0,75)}}{1+e^{(1,7)(1)(1-0,75)}} = \frac{e^{0,425}}{1+e^{0,425}} = \frac{1,53}{1+1,53} = 0,6049$$

$$\mathbf{P_i(1)} = 0,6049$$

A continuación se muestra una tabla donde aparecen todos los resultados de este ejercicio:

N	P(Q)				CCT
	ITEM 1	ITEM 2	ITEM 3	ITEM 4	
-3	0,0017	0,0000	0,0000	0,0000	0,0017
-2	0,0091	0,0004	0,0000	0,0000	0,0095
-1	0,0481	0,0059	0,0000	0,0000	0,0540
0	0,2177	0,0719	0,0010	0,0000	0,2906
1	0,6049	0,5000	0,0319	0,0001	1,1369
2	0,8938	0,9280	0,5000	0,0138	2,3356
3	0,9788	0,9940	0,9680	0,5000	3,4408

El análisis de ítems es un proceso cuantitativo y cualitativo mediante el cual se establece la calidad de los ítems de un instrumento, en relación con los propósitos para los cuales fueron elaborados. Su realización implica un saber profundo sobre el objeto de evaluación, la población evaluada, los propósitos de la evaluación y se requiere, además, conocer debidamente las técnicas de procesamiento de datos para hacer una adecuada interpretación de los indicadores estadísticos disponibles. El proceso de análisis de ítems debe conducir a la toma de decisiones en relación con la inclusión, exclusión, o modificación de ítems, a partir de la identificación clara de las posibles problemáticas de los mismos.

A continuación se describe cada uno de los indicadores que comúnmente se utilizan en el procesamiento de datos de los ítems, cuando se pretende proveer información cuantitativa necesaria para realizar el análisis de los mismos y decidir si se incluyen o no en una prueba:

Antes de explicar cada uno de los parámetros o indicadores que suelen utilizarse en el proceso de análisis de los ítems y de una prueba, definamos qué entendemos por parámetro en este caso.

Para nosotros un parámetro es un valor estadístico que refleja una cualidad del ítem y de la prueba. A continuación explicamos entonces a qué parámetros o indicadores nos estamos refiriendo.

- **DIFICULTAD**

  - **Definición:** indica la posición de la curva del ítem a lo largo de la escala de habilidad; entre más difícil es un ítem su curva estará localizada más a la derecha en la escala de habilidad.
  - **Justificación de uso**: es uno de los parámetros fundamentales en los Modelos de de la TRI. Es indicador base para la conformación de pruebas y de bancos de ítems, así como para establecer comparabilidad de escalas. Se requiere para obtener otros indicadores de ítems (curvas características, función de información).
  - **Interpretación:** los valores de dificultad oscilan entre menos infinito y más infinito en la escala logit, aunque en términos prácticos los ítems asumen valores entre –3.5 y +3.5, cuando el promedio de dificultades del grupo de ítems se centra en cero. Valores positivos y altos indican alta dificultad y los valores negativos indican baja dificultad.
  - **Criterio de aceptación:** regularmente se analiza la distribución de valores de dificultad del instrumento en relación con los valores de habilidad de la población evaluada para conceptuar sobre lo apropiado de la medición de dicho instrumento, de acuerdo con los propósitos que lo inspiraron. Un aspecto importante de análisis está dado por la densidad de ítems en un punto de la escala de habilidad en particular; así, se espera que no haya más de dos ítems de un mismo componente o contenido que midan con la misma dificultad. No se establecen de antemano valores de rechazo para este indicador; no obstante, ítems que sean respondidos correctamente por la totalidad de la muestra o que no sean respondidos correctamente por ninguno de los evaluados serán objeto de un reporte especial.

- **DISCRIMINACIÓN**

  - **Definición:** corresponde al "poder de un ítem para diferenciar a los evaluados en distintos niveles de habilidad frente a un constructo medido"[6] . Grado en el cual las respuestas a un ítem varían en relación con el nivel de habilidad. Se conoce también como la pendiente de la curva en el punto de máxima inflexión.
  - **Justificación de uso**: es, junto con la dificultad, parámetro fundamental de los ítems dentro del Modelo de Dos Parámetros. Indica en qué grado el ítem es respondido correctamente por las personas de alta habilidad e incorrectamente por las personas de baja habilidad.
  - **Interpretación:** los valores de discriminación oscilan, teóricamente, entre menos infinito y más infinito, aunque, en la práctica, los ítems presentan valores de discriminación entre 0 y +2. Valores que se aproximan a más infinito se corresponden con un patrón de Guttman (discriminación perfecta).
  - **Criterio de aceptación:** son aceptables los ítems con valores de discriminación superiores o iguales a 0.7. No obstante, en caso de tener un mayor número de ítems que los necesarios, la aceptabilidad se haría en orden del valor de la discriminación.

- **CORRELACIÓN PRODUCTO MOMENTO PUNTO MEDIDA**

  - **Definición:** relación entre la respuesta correcta a un ítem de una prueba y el valor de habilidad obtenido en dicha prueba. Este indicador supera las dificultades de la correlación punto biserial tradicional en tanto no es afectada por valores missing. Se calculará para la clave así como para las demás opciones de cada ítem.
  - **Justificación de uso**: permite inferir validez de los ítems en cuanto éstos se comporten como partes del instrumento.
  - **Interpretación:** puede tomar valores entre –1 y 1. Los valores positivos indican que la respuesta correcta al ítem está asociada a altos puntajes en la prueba; valores negativos indican que dicha asociación se da de manera inversa, es decir, altos puntajes en la prueba se asocian a una respuesta incorrecta al ítem.
  - **Criterio de aceptación:** son aceptables ítems con valores superiores a 0.25. Es deseable que la distribución de valores de correlaciones biseriales de los ítems de la prueba incluya valores superiores al criterio de aceptación. Igualmente, se debe seguir el orden del valor de la correlación punto biserial para la aceptación de los ítems.

6 ETS. (2000). ETS Standards for quality and fairness. Educational Testing Service. Princeton: New Jersey.

- **FUNCIONAMIENTO DIFERENCIAL DE ÍTEMS - DIF** [7]

  - **Definición:** grado en el cual un ítem presenta propiedades estadísticas diferentes en distintos grupos poblacionales, cuando se controla la habilidad de los grupos.
  - **Justificación de uso**: es un indicador de equidad en cuanto permite reconocer ítems que presenten comportamiento estadístico diferente en los distintos países participantes en el estudio, bien sea para proceder a ajustar dichos ítems –si se encuentra que el DIF se debe a fallas de construcción- o para reconocer necesidades particulares de ajuste o cualificación de procesos educativos –si se descartan problemas de construcción del ítem-. *En cualquier caso, debe garantizarse que la calificación final en el estudio se realice a partir de ítems que no presenten DIF.*
  - **Interpretación:** se interpreta el test t de significancia estadística para aceptación o rechazo de la hipótesis de funcionamiento diferencial que es equivalente al test de significancia Mantel – Haenszel.
  - **Criterio de aceptación:** se rechaza el funcionamiento diferencial cuando se encuentran valores inferiores a 1.96 de la prueba t con (n_1+n_2-2) grados de libertad.
  - **Cálculo**: el análisis de DIF se realiza con todos los ítems (o todos los grupos poblacionales), excepto con el que está siendo analizado, anclando sus dificultades (habilidades) a los valores de las dificultades (habilidades) para la totalidad de los ítems (o grupos poblacionales) incluyendo el que está siendo analizado. Se calcula un DIF de contraste que corresponde a la diferencia entre los tamaños de DIF de los ítems (o grupos poblacionales) en estudio, que es equivalente a la medida de DIF con procedimiento Mantel-Haenszel.

- **CURVA CARACTERÍSTICA DEL ITEM – ICC** [8]

  - **Definición:** función matemática que relaciona la probabilidad de éxito en un ítem con la habilidad medida por el ítem. Se debe calcular, para cada ítem, la ICC y la Curva Empírica del Ítem [9], tanto para la clave como para las demás opciones de respuesta. Se presentarán cuatro gráficas por prueba, por grado, en el Informe Internacional, solamente.
  - **Justificación de uso**: permite verificar el grado en el cual el modelo da cuenta de los resultados de la evaluación[10].
  - **Interpretación:** la curva característica del ítem para el Modelo de Dos Parámetros toma la forma de una ojiva logística. La interpretación consiste en verificar la correspondencia entre la curva empírica del ítem y la curva característica.
  - **Criterio de aceptación:** curvas empíricas (ECC) que se comportan de acuerdo con la curva característica del modelo.

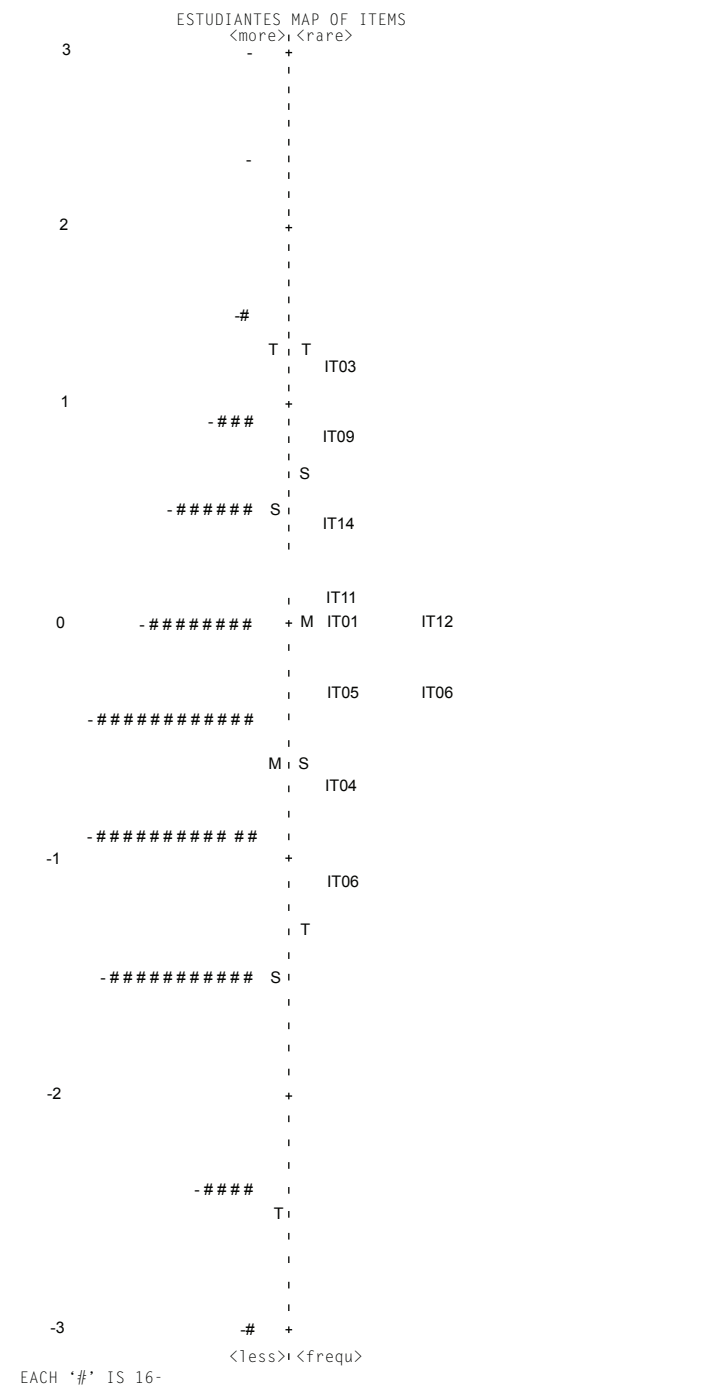
7 DIF, en inglés: Differential Item Functioning.
8 Item Caracteristic Curve
9 La curva basada en los puntajes observados se puede llamar Función Empírica de Respuesta al Ítem o Curva Empírica del Item (en inglés: Empirical Caracteristic Curve, ECC).
10 Se sugiere obtener las ICC para el modelo de dos parámetros y para el modelo de Rasch con el propósito de establecer el mejor ajuste de los datos.

- **AJUSTE PRÓXIMO Y LEJANO (INFIT Y OUTFIT)**
  - **Definición:** indica la correspondencia entre un grupo de datos y el modelo estadístico utilizado para representarlos. El ajuste próximo (infit) se refiere a la relación entre los datos que se encuentran cerca del valor de dificultad del ítem y el valor de dificultad; el ajuste lejano se refiere a la relación de los datos que se encuentran lejos de dicho valor de dificultad y esa dificultad. Ambos indicadores deben ser calculados para la clave. Para las demás opciones de respuesta debe calcularse el ajuste lejano (outfit).
  - **Justificación de uso**: la utilización de un modelo para representar datos debe fundamentarse en la verificación de que dicho modelo en verdad representa el comportamiento de los datos y, por ende, puede inferirse el cumplimiento de los supuestos de dicho modelo para los datos analizados.
  - **Interpretación:** los valores posibles se encuentran entre cero (0) e infinito positivo. El valor que determina el ajuste perfecto entre los datos y el modelo es 1. Los valores muy inferiores a 1 indican dependencia de los datos (paradoja de atenuación); valores superiores a 1 indican ruido en la información; valores superiores a 2 indican que el ruido es mayor que la información útil.
  - **Criterio de aceptación:** valores de ajuste entre 0.8 y 1.2.
  - **Cálculo:** se basa en la suma de los cuadrados de los residuos estandarizados. Esta suma se aproxima a una distribución chi cuadrado. Dividiendo esta suma por sus grados de libertad se obtiene un valor esperado de 1 y rango entre cero e infinito.

- **PROMEDIO DE HABILIDAD POR OPCIÓN**
  - **Definición:** promedio de las habilidades de quienes eligen cada opción de respuesta de un ítem. Se debe incluir, adicionalmente, la proporción de respuestas para cada opción. Esta información se presenta para el informe internacional de la aplicación piloto solamente e incluye los datos de todos los evaluados.
  - **Justificación de uso**: permite reconocer la proporción de personas que seleccionan una opción como respuesta correcta y su habilidad promedio. La combinación de estos dos datos proporciona información útil para determinar la validez de cada ítem en relación con los marcos conceptuales de las pruebas y valorar la calidad de cada opción.
  - **Interpretación:** los valores posibles de habilidad promedio se encuentran entre infinito negativo hasta infinito positivo. Se espera que el promedio de habilidad para la opción considerada clave sea el mayor de todos. El mapa de distribuciones de los promedios de habilidad de las opciones de todos los ítems indica el funcionamiento relativo de todas las opciones y puede ser interpretado con base en los marcos conceptuales de las pruebas. El orden de los promedios de habilidades permite comprender las complejidades de las opciones para la población evaluada. Las proporciones de respuesta para cada opción dan una idea de su atracción en una población particular.
  - **Criterio de aceptación:** el mayor promedio de habilidad debe ser el de la clave.

- **MAPA DE DISTRIBUCIÓN DE HABILIDADES Y DIFICULTADES**
  - **Definición:** Relación gráfica, en una misma escala, entre la distribución de las dificultades de los ítems y la distribución de las habilidades de los evaluados. Esta gráfica se presenta sólo en el informe internacional. Aunque no es un indicador estadístico en sí mismo, utiliza indicadores de personas y de ítems que ofrecen información pertinente para el análisis de los bloques y de los ítems que los conforman.

Un ejemplo de dicho mapa (salida gráfica del software winsteps)



- **Justificación de uso**: permite contrastar la dificultad de los ítems en una población particular y determinar si éstos se ajustan a la población; igualmente, es posible observar el cubrimiento de las habilidades por parte de los ítems. Permite identificar grupos de ítems o de personas que por su nivel de dificultad o habilidad, respectivamente, merezcan una atención especial en el análisis. Pueden ser contrastadas las expectativas de los constructores de ítems, en relación con el nivel de dificultad de los ítems, así como si la muestra seleccionada se comporta de acuerdo con los propósitos del diseño muestral.
- **Interpretación:** las habilidades y las dificultades se presentan en una escala que oscila entre menos infinito y más infinito. Si la distribución de habilidades tiene valores inferiores a la distribución de dificultades, quiere decir que para ese grupo poblacional los ítems resultaron difíciles. Por el contrario, si la distribución de habilidades tiene valores superiores a la distribución de dificultades, significa que para ese grupo poblacional los ítems resultaron fáciles.
- **Criterio de aceptación:** se espera que las distribuciones de habilidades y dificultades tengan posición y dispersión semejantes entre sí. De esta manera se entiende que el grupo de preguntas analizado cubre la totalidad de las habilidades de la población. Si las distribuciones no son semejantes, las diferencias deben ser interpretadas a la luz de los marcos teóricos de las pruebas y del propósito del estudio SERCE.

• **CORRELACIONES INTER – ÍTEM**

- **Definición:** correlaciones entre los ítems de un mismo bloque. Este indicador se presenta sólo en el Informe Internacional.
- **Justificación de uso**: indica el grado de relación entre dos ítems de un mismo bloque indicando si miden lo mismo (dimensionalidad del constructo). Se utiliza en análisis de la confiabilidad de un grupo de ítems.
- **Interpretación:** Si el valor de las correlaciones es positivo y alto indica que los ítems miden el mismo objeto; si, por el contrario, los valores son negativos, indica que los ítems miden objetos diferentes.
- **Criterio de aceptación:** las correlaciones deben ser positivas y altas.

• **ERROR ESTÁNDAR DE MEDICIÓN**

- **Definición:** corresponde a la desviación estándar de una distribución imaginaria de errores que representan la posible distribución de valores observados alrededor del valor teórico verdadero. Es un indicador de la confiabilidad.
- **Justificación de uso**: si se calcula el error de medición de cada habilidad estimada o dificultad estimada (o conjunto de habilidades o dificultades), se conoce la precisión de la medida o la estimación, orientando la toma de decisiones para la depuración de bases de datos y para el análisis de ítems.
- **Interpretación:** como en cualquier proceso de medición, se espera que el error sea cercano a cero. Errores demasiado grandes restan confianza en las estimaciones del parámetro. La escala de valores en la cual se reportan los

errores está asociada a la escala de medición utilizada, por lo cual no es posible establecer, de manera universal y de antemano, un valor mínimo aceptable de error.

- **Criterio de aceptación:** por lo general, diferentes autores coinciden en que la decisión de cuál sería un valor de error aceptable y cuál no, debe derivarse de un juicio profesional experto de quienes conocen procedimientos psicométricos, el instrumento de medida y su marco de fundamentación[11].

INDICADORES ESTADÍSTICOS PARA EL ANÁLISIS DE GRUPOS DE ÍTEMS

En consideración a que el diseño del estudio contempla que los instrumentos (pruebas) estén conformados de acuerdo con agrupaciones de ítems en bloques, aplicados de manera sistemática en cuadernillos editados con arreglos distintos de dichos bloques, el análisis de ítems debe incluir una fase de análisis de indicadores estadísticos que den cuenta del comportamiento de tales agrupaciones.

En tal sentido, a continuación se describen los indicadores propuestos para el análisis de bloques; es de anotar que dado que la aceptación o rechazo de ítems se hace con base en sus indicadores individuales, para los indicadores de grupos de ítems no se define un criterio de aceptación.

- **PROMEDIO**

Se requiere obtener: Promedio de Habilidades, Promedio de Dificultades y Porcentaje de Respuestas Correctas.

- **Definición:** promedio de las dificultades de los ítems del bloque y de las habilidades de las personas que abordan el bloque.
- **Justificación de uso**: permite conocer el comportamiento del bloque en diferentes grupos poblacionales, en relación con la posición de dicho bloque en distintos cuadernillos. También permite observar el comportamiento relativo de los distintos grupos poblacionales, de acuerdo con el diseño muestral.
- **Interpretación:** si el bloque se encuentra ajustado a la población, el promedio de habilidades debe ser aproximadamente igual al promedio de dificultades. Si el promedio de habilidades es superior, significa que el bloque fue relativamente fácil para ese grupo poblacional; si por el contrario el promedio de habilidades es inferior al promedio de dificultades, significa que el bloque fue relativamente difícil para ese grupo poblacional. El porcentaje de respuestas correctas aporta al análisis intrabloque en cuanto constituye un indicador general de la manera en que los evaluados abordaron cada bloque.

- **DESVIACIÓN ESTÁNDAR**

Se requiere obtener la desviación estándar de habilidades y de dificultades.

- **Definición:** medida de la dispersión de la distribución de las dificultades de los ítems de un bloque y de la distribución de las habilidades de las personas que abordan dicho bloque.

11 AERA, APA, NCME. (1999). Standards for educational and psychological testing. Washington: AERA.

- **Justificación de uso**: permite valorar la homogeneidad/heterogeneidad de los valores de dificultad del grupo de ítems de un bloque, así como de los valores de habilidad del grupo poblacional que aborda dicho bloque.
- **Interpretación:** desviaciones estándar altas (superiores a 1 en valores logit) indican heterogeneidad del grupo de datos; desviaciones estándar bajas (inferiores a 1 en valores logit) indican homogeneidad del grupo de datos. Esta información se contrasta con el respectivo valor del promedio para comprender el desempeño de una población particular.

• **PUNTUACIÓN MÁXIMA Y PUNTUACIÓN MÍNIMA**

- **Definición:** puntuaciones logit más alta y más baja alcanzadas por una población particular.
- **Justificación de uso**: permite reconocer los extremos de la distribución de habilidades en una población particular y contrastarlos con las expectativas de los evaluadores.
- **Interpretación:** si el valor máximo en logits alcanzado por una población es inferior al valor de habilidad equivalente a n-1 ítems correctos, indica que el grupo de ítems resultó complejo para las personas de más alta habilidad. Si el valor mínimo, en logits, alcanzado por una población es superior al valor de habilidad equivalente a 1 ítem correcto, indica que el grupo de ítems resultó fácil para las personas de más baja habilidad. Se espera que los valores mínimos y máximos correspondan a 1 respuesta correcta y a n-1 respuestas correctas, respectivamente.

• **CONFIABILIDAD**

- **Definición:** indica el grado de precisión en la medición. En la TRI se utiliza la Función de Información como indicador de la confiabilidad.
- **Justificación de uso**: es importante conocer la precisión de la medición en cada punto de la escala de habilidad, para cada uno de los bloques, en las distintas poblaciones.
- **Interpretación:** a mayor valor de la función de información es mayor la precisión en la medida.

SOFTWARE PARA EL PROCESAMIENTO DE LOS DATOS

El procesamiento de los datos de una prueba debe ofrecer información suficiente para efectuar el análisis de ítems con los indicadores sugeridos anteriormente.

En el mercado se cuenta con oferta considerable de software que opera con los supuestos de la Teoría de Respuesta al Ítem y que implementa las funciones matemáticas de sus diferentes modelos, enunciados en el epígrafe anterior de este texto. La diferencia entre un software y otro, ya sea que estén o no inspirados en un mismo modelo, radica principalmente en aspectos como el tamaño poblacional y la longitud de los instrumentos que les es posible procesar, en la formulación matemática particular que implementan y, quizás lo más importante, desde el punto de vista de un proceso de evaluación, en la confiabilidad (precisión) de los datos que arroja.

Se revisamos las distintas características técnicas (la cantidad de indicadores que reporta el software y la precisión en la estimaciones de los mismos; la cantidad de datos que tiene capacidad de analizar; los requerimientos de conformación de bases de datos; la convergencia en las estimaciones y la interfase gráfica) de programas de software disponibles en el mercado, tales como Bilog MG, Multilog, Parscale, Rascal y Winsteps se puede llegar a la conclusión que este último ofrece las mayores ventajas para el procesamiento de datos de una prueba.

Winsteps es un software que opera en plataforma windows y que implementa los principios de la Teoría de Respuesta al Ítem para construir mediciones objetivas a partir de una base de datos sencilla en la cual se especifican personas (evaluados) y sus respuestas a un grupo de ítems.

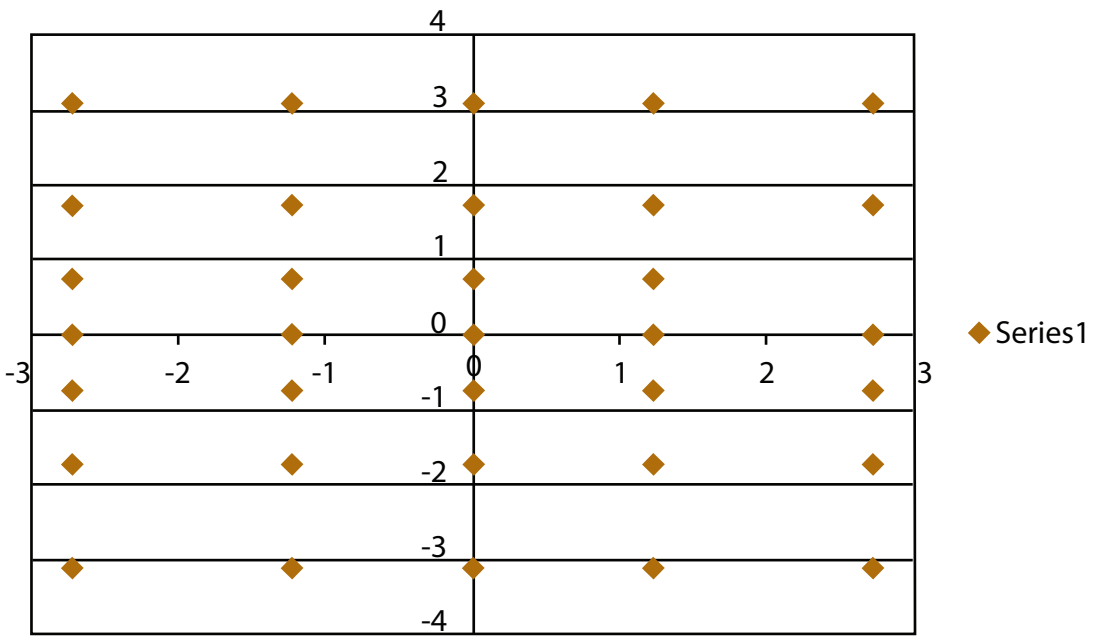
Puede trabajar de manera simultánea con varios formatos de ítems (dicótomos, de selección múltiple con única y con múltiple respuesta y de crédito parcial); ofrece gran variedad de reportes tabulares y con interfases gráficas, en los cuales se presenta de manera detallada y también resumida, el comportamiento de las poblaciones y de los ítems analizados. Una característica importante es que el software señala con claridad los datos que presentan comportamientos por fuera de lo esperado. Los datos missing no son un problema para las estimaciones que realiza este software.

La generación de escalas de calificación usando Winsteps constituye un proceso relativamente simple toda vez que el software permite el procesamiento de datos, agrupados según variables de interés, sin requerir modificaciones en la estructura de la base de datos original. Winsteps también permite prefijar valores (anclar) de los parámetros para facilitar procesos de comparabilidad (equating).

Winsteps puede procesar hasta 10.000.000 de personas y 30.000 ítems y cada ítem puede contemplar hasta 255 categorías de calificación (el de mayor capacidad, para ítems y categorías, en el mercado). El software tiene un manual detallado para orientar su uso y la interpretación de la información que arroja; la firma Winsteps ofrece apoyo técnico en línea, oportuno y eficaz, para solventar inquietudes tanto de procesamiento como de interpretación de datos.

Ejemplos de algunas salidas gráficas de este software

Para comprobar gráficamente la unidimensionalidad, utilizando el winsteps.

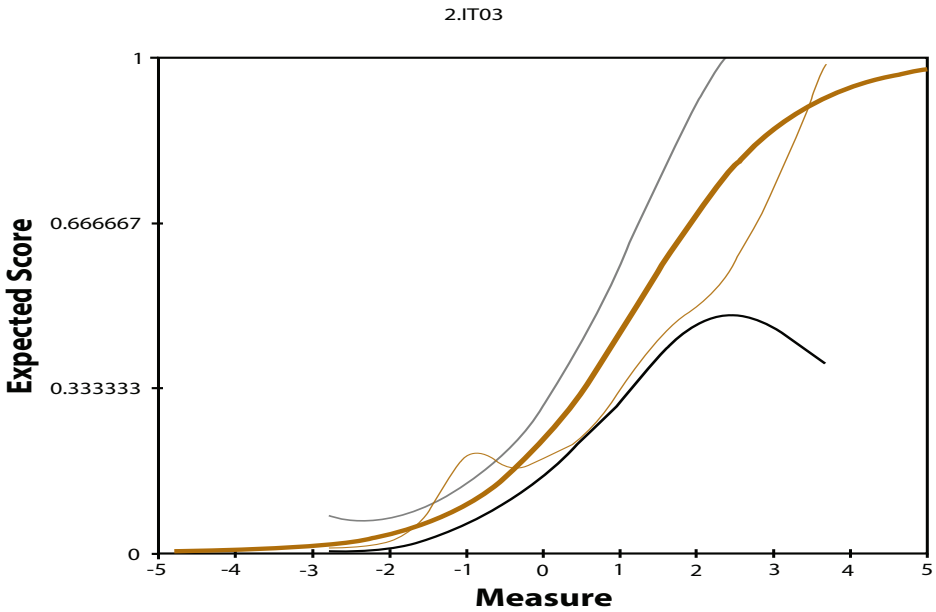


En este caso la interpretación del gráfico consiste en que "no se cumple el supuesto de unidimensionalidad pues cada conjunto de ítems está midiendo un rasgo latente diferente".

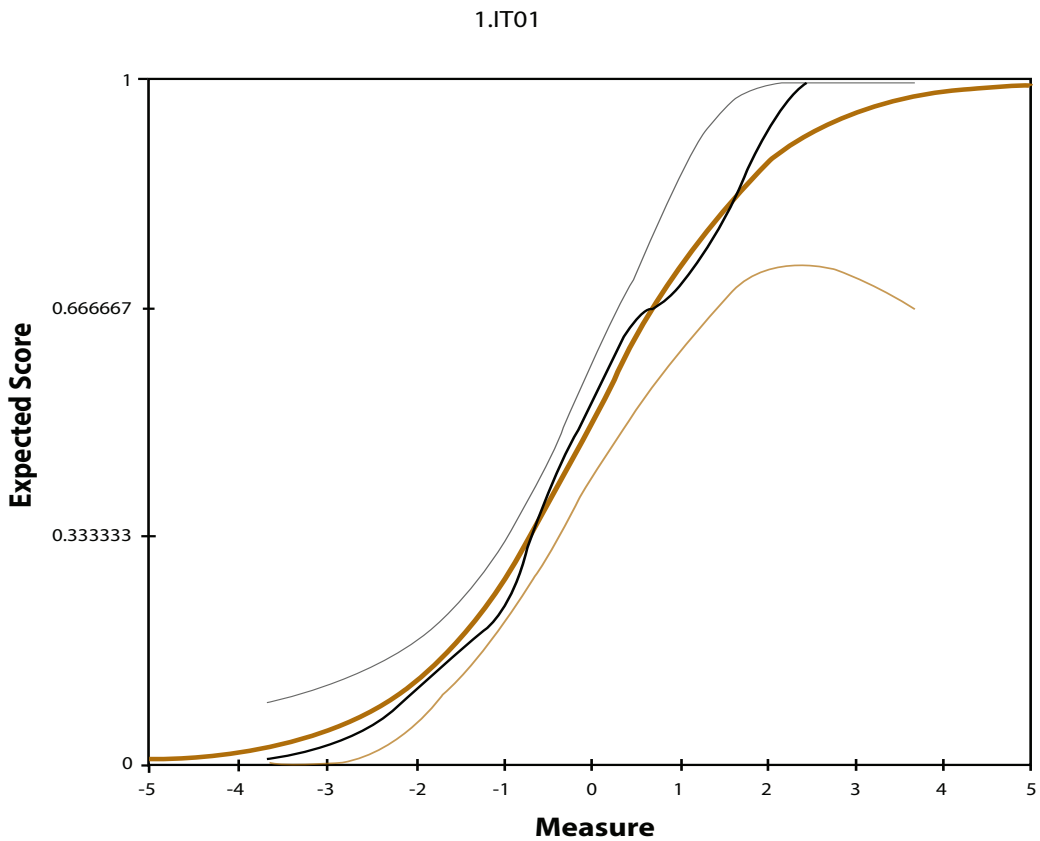
Comprobación gráfica del supuesto de la curva característica del ítem

Cada modelo tiene una curva característico, la curva ideal (aparece con color rojo en el gráfico que aparece a continuación). El elemento que se necesita contrastar en una medición es la curva real del ítem (aparece en color azul) contra el modelo previsto teórico. Hay un grado de diferencia entre el comportamiento real de las personas y la curva ideal del modelo.

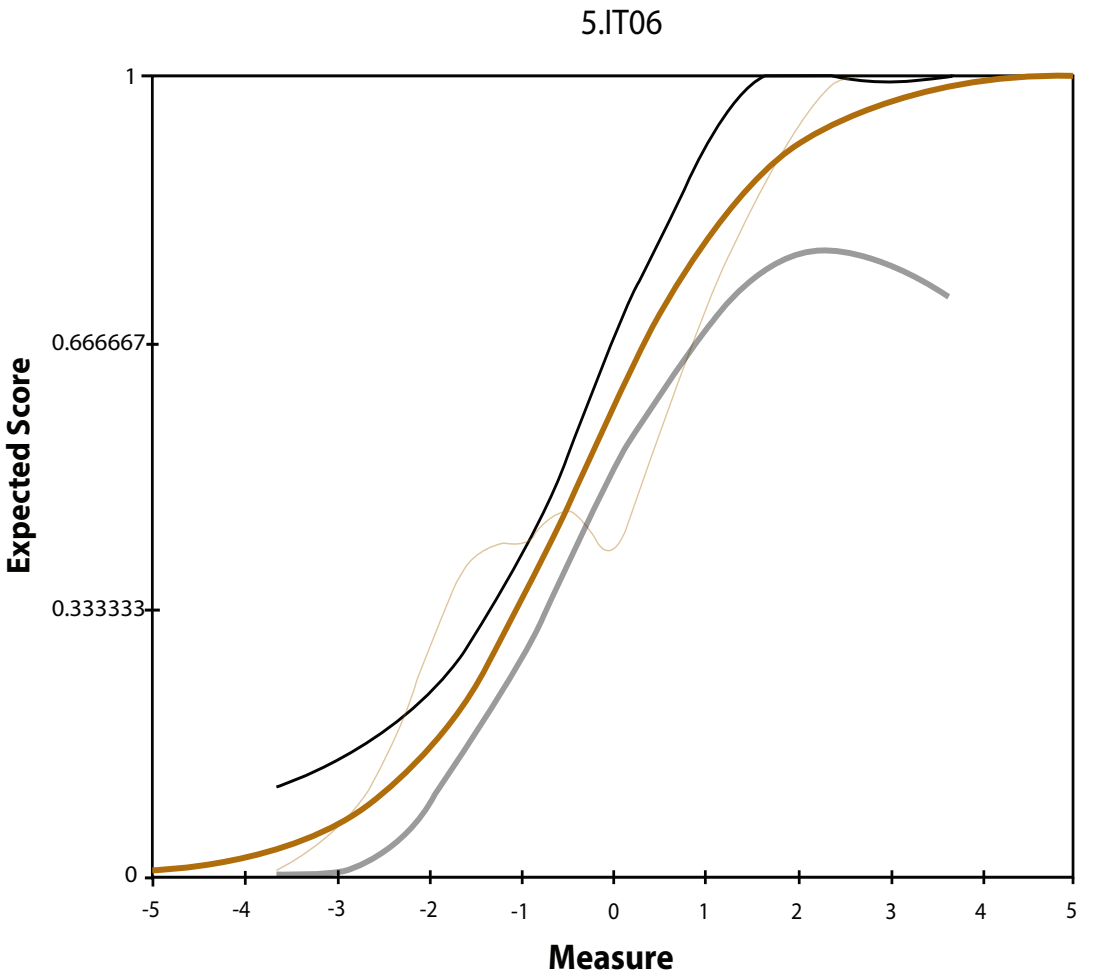
Las líneas grises representan los niveles extremos. Se espera que la azul si es perfecta se superponga sobre la roja y que en ningún caso se salga fuera de las curvas extremas.



En el caso del ítem representado en el gráfico anterior, solo se sale de la franja para estudiantes de habilidad entre – 2 y -1.



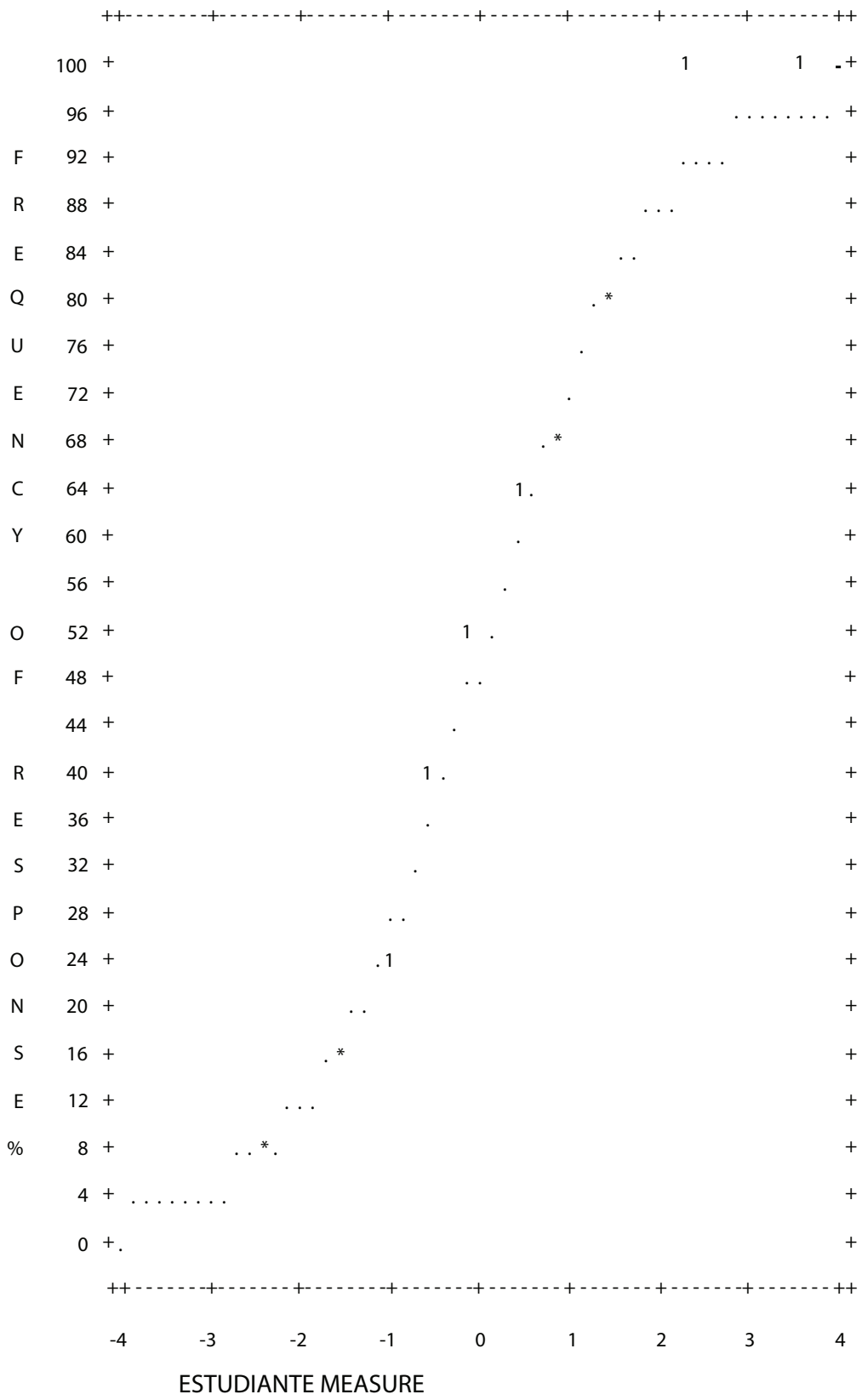
En el caso del ítem representado en el gráfico anterior, este se ajusta totalmente al modelo ideal, pues se mantiene dentro de la franja marcada por las curvas grises. Ello significa que el mismo cumple el supuesto de la curva característica del ítem.



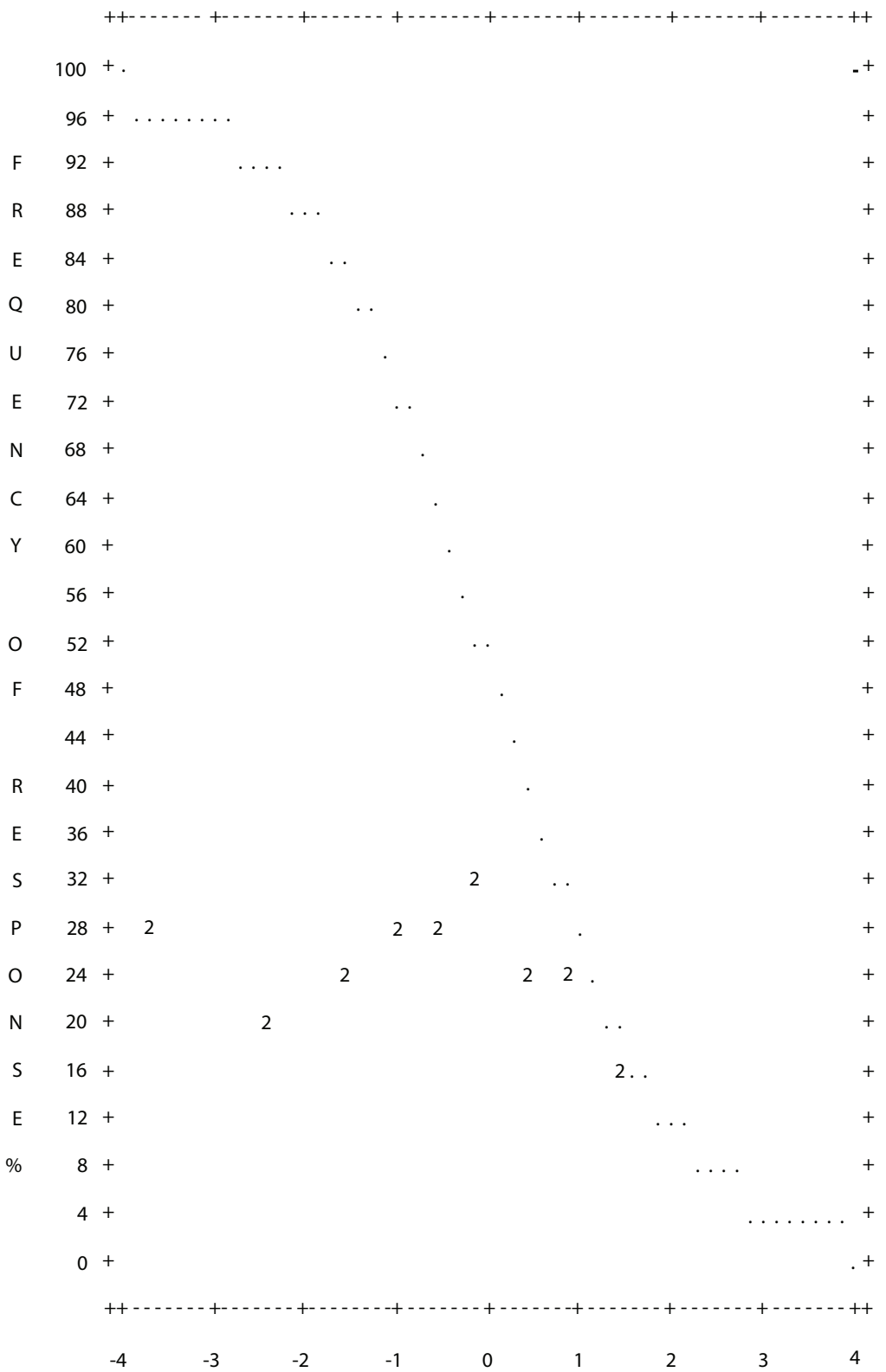
Resulta evidente que el ítem representado en el gráfico anterior, no cumple con el supuesto de la CCI para estudiantes de niveles intermedio de habilidad.

Cada distractor tiene también su curva característica. A continuación vemos un ejemplo:

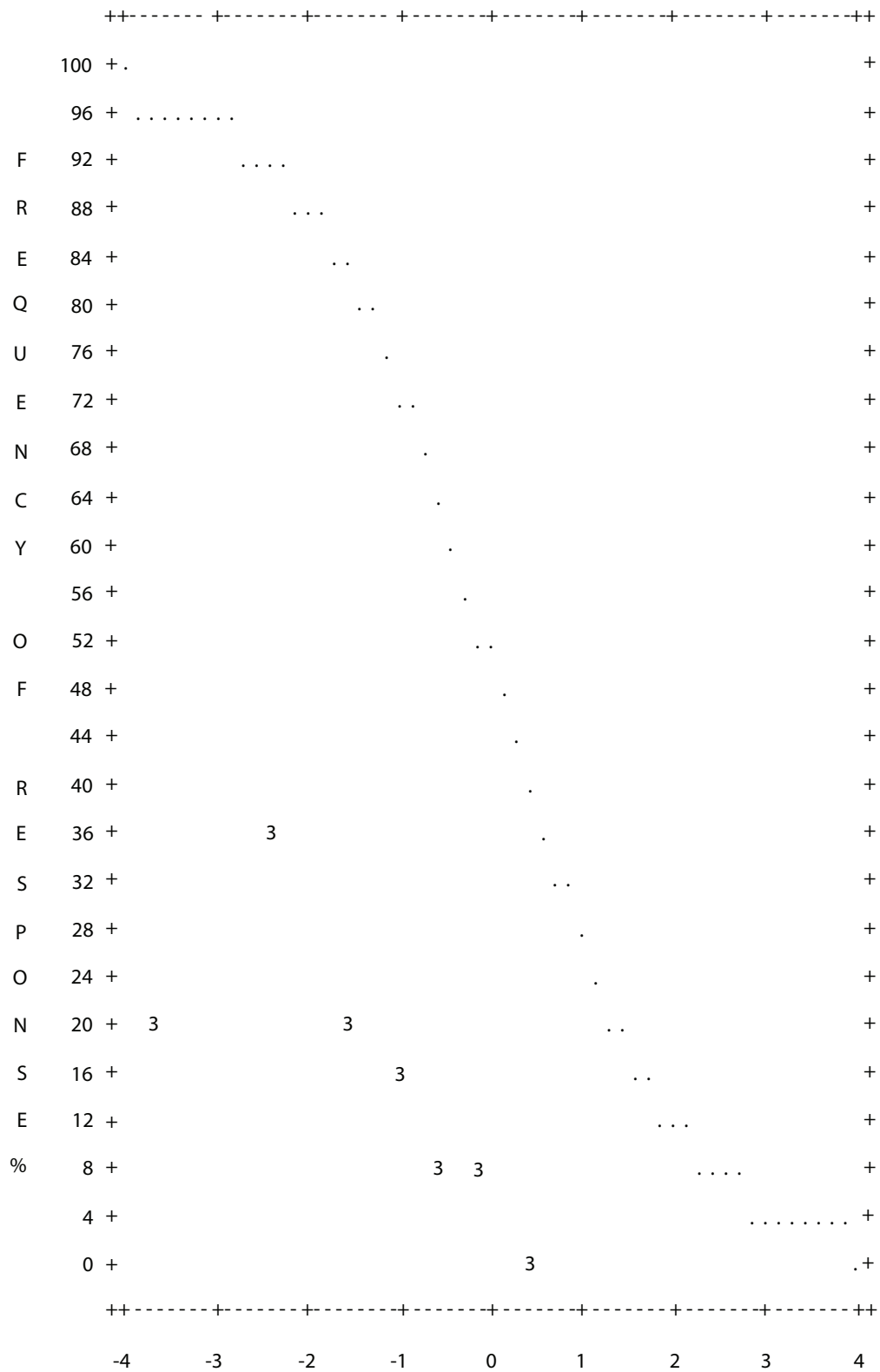
EMPIRICAL CODE FREQUENCIES: "1" : 1. IT01

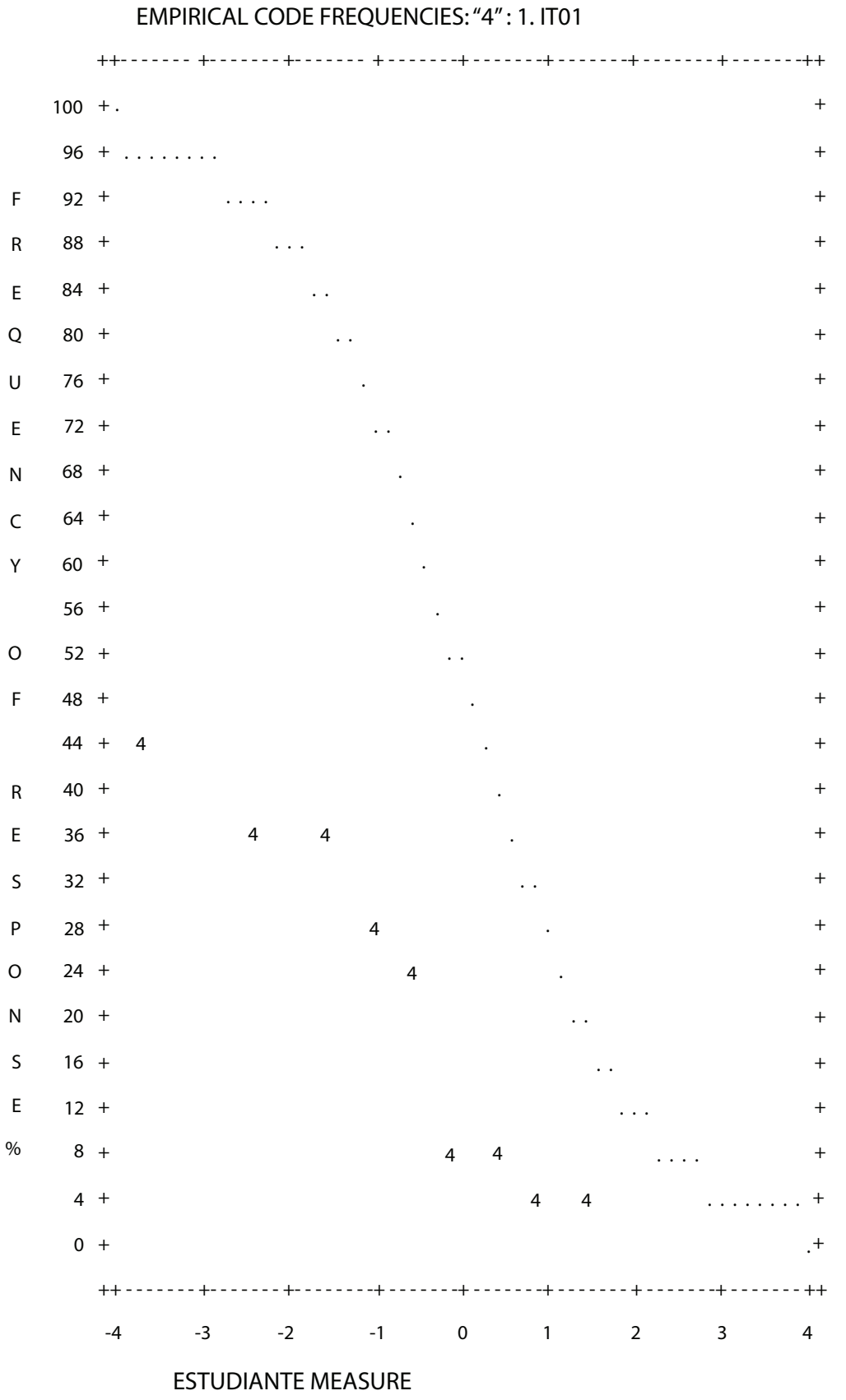


EMPIRICAL CODE FREQUENCIES: "2" : 1. IT01



EMPIRICAL CODE FREQUENCIES: "3" : 1. IT01




EMPIRICAL CODE FREQUENCIES: "4" : 1. IT01
FREQUENCY OF RESPONSE %
100
96
92
88
84
80
76
72
68
64
60
56
52
48
44
40
36
32
28
24
20
16
12
8
4
0
-4
-3
-2
-1
0
1
2
3
4
ESTUDIANTE MEASURE

BIBLIOGRAFÍA

- Badger, E. y Thomas, B. 1992. *Open ended questions in reading*. Washington, ERIC Clearinghouse on Tests Measurement and Evaluation.
- Cohen, A. y Woollack, J. 2004. Helpful tips for creating reliable and valid classroom test *Handbook on Test Development*. U. Wisconsin.
- Cheung, D. y Bucat, R. 2002. *How can we construct good multiple choice ítems?* Hong Kong, Sciens and Technology Education Conference.
- Fenton, N. E. y Pfleeger, S. L. 1997. *Software metrics. A rigurous and practical approach.* Boston, PWS Pub.
- Haladyna y Downing. 1989. A taxonomy of multiple choice ítem writing rules *Apply Measurement in Education*. Vol. 1.
- Haladyna, T. 1994. *Development and validatin multiple choice test ítems*. New Jersey, Lawrence Earlbaum Associates.
- Hambleton, R. y Zaal, J. 1994. *Advances in educational psychological testing*. Boston, Kluwer Academic Publishers.
- ICFES. 2004. *Estándares para la construcción de pruebas.* Grupo de Evaluación de la Educación Superior. Bogotá, ICFES.
- Martínez Árias, María Rosario y otras. Psicometría. Alianza Editorial S.A, 2006, Madrid, España.
- Messick, S. 1989. Validity. R.L. Linn (Ed.). *Educational measurement* New York, Macmillan, 3a ed., pp. 13-103.
- Roberts, D. 1993. An empirical studying on the nature of trick questions. *Journal of educational measurement*. Vol. 30.
- B. Baker, Frank. Fundamentos de la Teoría de Respuesta al ítem. Universidad de Wisconsin.
- B. Baker, Frank. A criticism of Scheuneman's items bias techniques. Journal of Educational Measurement, 1981.
- Lord, F.M. Ateory of test scores. Psychometric Monograph, num. 7, 1952.
- Muniz Fernández, José. Teoría de Respuesta a los ítems: Un nuevo enfoque en la evolución psicológica y educativa. Ediciones Pirâmide, S.A, Madrid, 1990.
- Seminário Regional de evaluación de la educación, zona occidente. Taller sobre elaboración de ítems, ICFES, CAli, Colômbia, 2006.


ideice