Inv Ed Med 2014;**3(9)**:40-55

Investigación en Educación Médica

http://riem.facmed.unam.mx

METODOLOGÍA DE INVESTIGACIÓN EN EDUCACIÓN MÉDICA

Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas

Iwin Leenen

Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., México

Recepción 26 de septiembre de 2013; aceptación 30 de octubre de 2013

PALABRAS CLAVE
Teoría clásica de los tests; teoría de respuesta al ítem; psicometría; evaluación educativa; análisis de ítems; México.

Resumen
La teoría clásica de los tests (TCT) y la teoría de respuesta al ítem (TRI) constituyen los dos enfoques principales de la psicometría. Aunque los fundamentos de la TRI se elaboraron a mediados del siglo XX y numerosas publicaciones han argumentado la superioridad teórica de la TRI sobre la TCT, el enfoque clásico sigue siendo, por mucho, lo más utilizado para la evaluación educativa, también en el campo de la educación médica. En este artículo, se revisan los fundamentos y conceptos centrales de ambos enfoques psicométricos y se esbozan las posibles ventajas de los modelos de la TRI en el contexto de la evaluación educativa en las ciencias de la salud. Sin embargo, al evaluar los supuestos que subyacen los modelos TRI básicos, es notable una discrepancia significativa entre los mismos y la compleja realidad en la evaluación educativa. Dicha discrepancia lleva a la conclusión que, para poder aprovechar las ventajas de la TRI, muchas veces es necesario considerar modelos más complejos que los conocidos tradicionalmente, como los modelos multidimensionales y/o modelos que toman en cuenta dependencias entre preguntas particulares.

KEYWORDS
Classical test theory; item response theory; psychometrics; educational measurement; item analysis; Mexico.

Virtues and limitations of item response theory for educational assessment in the medical sciences

Abstract
Classical test theory (CTT) and item response theory (IRT) constitute the two main paradigms in psychometrics. Although the foundations of IRT were already introduced in the middle of the twentieth century and despite the numerous publications since which show the theoretical superiority of IRT over CTT, the classical approach is still, by far, the most commonly used

Correspondencia: Secretaría de Educación Médica, Facultad de Medicina, Universidad Nacional Autónoma de México. Edif. B, 3er piso, Av. Universidad N° 3000, Circuito escolar CU, C.P. 04510, México D.F., México. Teléfono: 5623 2300, ext. 43034. *Correo electrónico:* iwin.leenen@gmail.com

ISSN: 2007-5057 – see front matter

for educational measurement, not the least in the field of medical education. In this article, I revise the fundamentals and basic concepts of both psychometric approaches and highlight the advantages that IRT models may offer in the context of educational assessment in the health sciences. However, based on an evaluation of the assumptions underlying the most commonly used IRT models, it is argued that these assumptions are significantly discrepant with the complex reality often encountered in educational measurement. As a result, it is concluded that, in order to take proper advantage of the IRT framework, often more complex models, beyond the traditionally known, must be considered, including multidimensional models and/or models that take into account local dependencies among test items.

Introducción

A la luz del objetivo de formar profesionales de la salud competentes y preparados para proporcionar atención médica de calidad, se considera una tarea esencial de la educación médica monitorear y evaluar de forma continua el proceso educativo de los estudiantes de medicina. En este sentido, la psicometría juega un papel importante dentro del campo de la educación médica, ya que esta disciplina investiga cómo medir y evaluar de forma óptima los constructos y atributos centrales en el aprendizaje de los estudiantes (como conocimientos, competencias, actitudes, entre otros). Por ejemplo, la psicometría permite analizar la validez de los instrumentos utilizados para la evaluación educativa y propicia el desarrollo de ideas o propuestas para mejorar dichos instrumentos.

Existen dos enfoques principales de la psicometría: la teoría clásica de los tests (TCT) y la teoría de respuesta al ítem (TRI). El primero, que se conoce también como *modelo de la puntuación verdadera* o *teoría del error de medición*, se cimentó en las ideas originales de Charles Spearman, cuyas elaboraciones matemáticas publicadas al inicio del siglo XX implicaban la diferenciación de los conceptos *puntuación verdadera* y *puntuación observada* como resultado de la aplicación de una prueba.[1,2] La TRI, por otro lado, cuyos fundamentos se elaboraron en la segunda mitad del siglo pasado a partir de las contribuciones seminales de Louis Guttman, Frederic Lord y George Rasch, aproxima el análisis de las respuestas en una prueba de forma radicalmente diferente, enfocándose en los componentes constituyentes de la misma (es decir, los ítems) en vez del resultado global de la medición.[3-5]

Gracias a los avances tecnológicos y los nuevos desarrollos teóricos, la TRI creció en relevancia e importancia durante las últimas tres décadas y cada vez más se considera una alternativa viable para la TCT. Actualmente, constituye una familia muy extensa de modelos psicométricos, los cuales tienen en común que relacionan formalmente —generalmente a través de una(s) ecuación(es) matemática(s)— las características latentes (es decir, hipotéticas, no observables) de los ítems en una prueba y de las personas que la contestan, con el fin de llegar a afirmaciones (probabilísticas) de la conducta de cada persona en cada ítem. Aunque entre los expertos en psicometría existe consenso general sobre la superioridad teórica de la TRI, el enfoque principal en contextos aplicados para el análisis de los resultados de los tests sigue siendo la TCT. Específicamente, en el área de la educación médica son escasos los estudios que analizan los datos de instrumentos de evaluación dentro del marco de la TRI.

Este artículo dará una introducción conceptual de los dos enfoques principales de la psicometría; espera dar una presentación clara de los conceptos claves de cada uno de estos paradigmas y quiere invitar a los investigadores en educación médica a considerar —y reflexionar críticamente sobre— la perspectiva que ofrece la TRI como alternativa para la TCT. Además, se espera aclarar que la TRI es mucho más que los dos o tres modelos que se suelen presentar en los artículos introductorios y que la familia de modelos TRI incluye miembros cuyos supuestos se ajustan mejor a los contextos típicos de evaluación en medicina. La introducción a los modelos en este artículo es necesariamente limitada; para un tratamiento más completo, el lector interesado puede consultar las diversas publicaciones que existen sobre el tema, tanto en español[6-8] como en inglés.[9-13]

La primera y segunda sección, revisan los conceptos y supuestos básicos y la lógica subyacente de la TCT y la TRI, respectivamente. En la tercera sección se comparan ambas aproximaciones y se evalúan las diferencias desde un punto de vista teórico a través de un análisis de (algunos de) los argumentos que manejan los expertos para colegir la superioridad de la TRI. La cuarta sección reconsidera los modelos más comunes de la TRI; en particular, se contrastan sus supuestos con la realidad compleja con que se suele topar en la evaluación educativa y se discuten algunos modelos alternativos que pueden ofrecer una respuesta a los inconvenientes percibidos. La última sección concluirá el artículo con unas reflexiones generales sobre las ideas expuestas.

Conceptos básicos de la teoría clásica de los tests

Puntuación verdadera y la ecuación básica de la TCT

La TCT es una teoría sobre la medición que se obtiene al aplicar un instrumento a una persona. Consideremos, en primera instancia, la aplicación del instrumento a sólo una persona, digamos la persona p, y representemos el resultado de esta medición como x_p (lo cual, entonces, corresponde a una puntuación codificada como un número real). Spearman[1,2] reconoció que, debido a la interferencia de factores perturbantes, x_p generalmente *no* coincide con la medición que uno realmente desea tener, es

decir, que el resultado observado va acompañado con un error de medición. Los factores perturbantes que causan el error de medición pueden tener su origen en la persona, en el instrumento, o en la situación. Como ejemplo del primer tipo de errores, se puede pensar en la medición de la presión arterial, que fluctúa considerablemente en el transcurso de un día por lo cual una única medición suele ser insuficiente. Para un ejemplo del segundo tipo de perturbaciones, considérese el termómetro, el cual intercambia calor con el cuerpo sujeto de la medición y, por lo tanto, no dará la temperatura exacta de este cuerpo. Factores perturbantes que tienen su origen en la situación ocurren, por ejemplo, en la aplicación de un examen mientras que en la plaza de a lado un candidato presidencial ha organizado un mitin y pronuncia su discurso electoral bajo los fuertes aplausos y las porras de sus simpatizantes. Dentro de la TCT, se formulan supuestos sobre el efecto de los factores perturbantes (es decir, sobre el error de medición) y se desarrollan procedimientos para cuantificar su influencia en el resultado obtenido.

El supuesto básico de la TCT es que, en cada medición, el error se extrae aleatoriamente de alguna distribución de probabilidad.[14] Si ε_p representa el error que acompaña la medición x_p, la diferencia

$$x_p - \varepsilon_p$$

corresponde con la "puntuación purificada", es decir, la puntuación de la cual se ha quitado el error de medición. La teoría clásica denomina este resultado *puntuación verdadera* y se representa por τ_p. Es decir, se define

$$\tau_p = x_p - \varepsilon_p,$$

lo cual es algebraicamente equivalente a

$$x_p = \tau_p + \varepsilon_p.$$

Mientras que x_p es el valor observado de la medición y por lo tanto conocido, la puntuación verdadera τ_p y el error de medición ε_p son constructos hipotéticos —que existen sólo gracias a la teoría— y desconocidos, o bien, *latentes*. (Nótese que en este artículo me adhiero a la costumbre de representar parámetros latentes por minúsculas griegas y los valores observados por minúsculas romanas; las mayúsculas se reservan para representar variables). Aunque nunca se conoce con exactitud la puntuación verdadera y el error asociados con una medición concreta, dentro de la TCT se han desarrollado métodos que permiten llegar a conclusiones sobre estas entidades a partir de los datos de una muestra.

Para aclarar y precisar las implicaciones del supuesto básico mencionado anteriormente de que el error de medición es el resultado de una extracción de alguna distribución de probabilidad, considérese el siguiente experimento mental. Supongamos que fuese posible repetir un gran número de veces la medición de la persona *p* bajo *circunstancias similares* a las de la medición inicial, es decir, sin que las aplicaciones anteriores influyesen en las nuevas mediciones (no hay efectos de memoria, fatiga, aburrimiento, etc., como si se le lavase el cerebro a la persona antes de cada nueva aplicación). Entonces, el supuesto básico implica que en cada una de estas repeticiones (a) se extrae un nuevo valor para el error de medición de su distribución probabilística, mientras que (b) la puntuación verdadera *no* cambia. En otras palabras, considerando las réplicas hipotéticas de la persona *p*, el error de medición es una variable aleatoria, la cual se representa por E_p, y la puntuación verdadera es una constante (τ_p). Esto implica que la puntuación observada de la persona *p* también es una variable aleatoria: X_p.

La **Figura 1** ilustra esta idea gráficamente. Para concretar el ejemplo, supongamos que las puntuaciones son de un examen clínico objetivo estructurado (ECOE) que se ha calificado como un porcentaje, es decir, son calificaciones sobre 100 (que se calcularon a partir de las calificaciones en una serie de estaciones). La distribución en la parte superior izquierda representa la distribución de probabilidad de E_p para una primera persona (p=1). Para este ejemplo se escogió una distribución normal (aunque el supuesto de normalidad no es parte del núcleo de la TCT). En la tabla debajo de la distribución de E_1 se resumen los resultados obtenidos en ocho de las réplicas hipotéticas que se realizaron en el experimento mental. Se observa que, por un lado, la puntuación verdadera de esta persona (τ_1 = 64.30) es una constante; el valor de E_1, por otro lado, es diferente en cada réplica y se ha extraído de la distribución arriba (por ejemplo, ε_{11}, el error que acompaña la primera medición de la primera persona, es igual a −2.31; ε_{12}, el error de su segunda medición, es +2.59, etc.). Puesto que el modelo supone que la puntuación observada es la suma de una constante y una variable, X_p varía también entre las réplicas de la misma persona.

La gráfica de la distribución en la **Figura 1** tácitamente refleja otro supuesto de la TCT: el valor esperado (es decir, la media) de la distribución de probabilidad que se supone para E_p es igual a 0, para cada persona *p*. En algunas réplicas la puntuación observada sobreestima la puntuación verdadera, en otras la subestima, pero, a la larga, los efectos positivos y negativos de los factores perturbantes se equilibran. Por otro lado, la varianza de la distribución de probabilidad de E_p, denotada σ^2_{Ep}, es un índice de la precisión de las mediciones de la persona *p*: si es grande, los valores de E_p fluctúan mucho entre las réplicas; en el caso extremo de que $\sigma^2_{Ep} = 0$, E_p siempre asume el mismo valor, igual a la media 0, y entonces no hay error. Nótese que de lo anterior directamente sigue que:

$$\mathcal{E}(X_p) = \tau_p \quad \text{y} \quad \sigma^2_{Xp} = \sigma^2_{Ep}.$$

Este resultado implica que la puntuación verdadera de la persona *p* se podría estimar a partir de la media de las puntuaciones observadas en una muestra de réplicas y que la varianza de las mismas puntuaciones observadas sería un indicador de la precisión de las mediciones para esta persona.

La parte superior derecha de la **Figura 1** representa el mismo proceso para otra persona (p=2). Nótese que la distribución de probabilidad de los errores de medición de esta persona tiene mayor varianza: la TCT, en su desarrollo inicial, no restringe que las distribuciones de probabilidad para diferentes personas sean idénticas.

En la realidad, sin embargo, varios obstáculos prácticos impiden obtener réplicas de la medición de la misma persona bajo las circunstancias especificadas en el experimento mental. Con el objetivo de poder estimar los parámetros de interés, la TCT cambia un poco el enfoque y añade unos nuevos supuestos al modelo: en lugar

de considerar a las personas por separado, se extiende la teoría a una *población* de personas. La tabla en la parte inferior de la **Figura 1** sirve para aclarar dicha extensión: la conceptualización de la TCT para una población sigue, tal como la construcción de la tabla, un proceso de dos pasos. Primero, para cada persona se saca independientemente un error de medición ε_p de su distribución, el cual, según lo anteriormente expuesto, se suma a la puntuación verdadera τ_p de esta persona para obtener la puntuación observada x_p. Segundo, se definen tres nuevas variables *E*, *T* y *X*, que representan la variación del error de medición y las puntuaciones verdaderas y observadas, respectivamente, dentro de la población de personas. Como muestran las tres columnas correspondientes de la tabla inferior de la **Figura 1**, hay variación en las tres variables (contrario al caso de las réplicas dentro de cada persona, donde únicamente varían el error de medición y la puntuación observada). Estas tres variables se relacionan en la ecuación central de la TCT:

$$X = T + E \qquad (1)$$

Confiabilidad y error estándar de medición

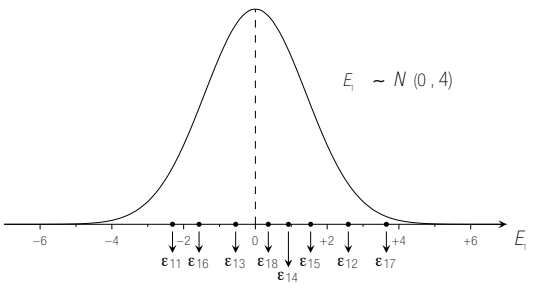
En la sección anterior se identificó la varianza $\sigma^2_{E_p}$ como un índice de la precisión de las mediciones de la persona *p*. Además, se mencionó que, en principio, las varianzas de distintas personas pueden diferir. Sin embargo, al considerar la población de personas y la variable *E*, cuyos valores se extraen de las respectivas distribuciones individuales (es decir, son valores realizados de las variables E_p asociadas con las distintas personas), la teoría clásica incorpora como supuesto adicional que dichas distribuciones individuales sean idénticas y, particularmente, que para cualquier persona *p* se cumpla

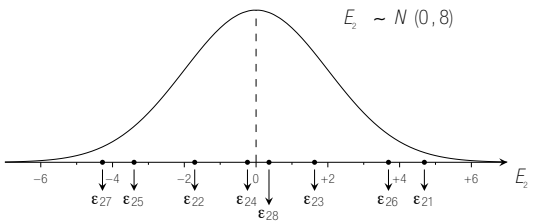
$$\sigma^2_{E_p} = \sigma^2_E. \qquad (2)$$

En relación con la **Figura 1**, este supuesto implica que (a) se cambien las distribuciones en la parte superior para que sean idénticas (con la misma varianza) y, por consiguiente, (b) que la varianza de los valores de E_1 en la segunda fila de la tabla de la persona 1 sea igual a la varianza de los valores en la fila de E_2 de la persona 2 y que, además, (c) las varianzas en estas filas sean iguales a la varianza de los valores en la última columna de la tabla inferior.

El desarrollo del modelo hasta el momento permite derivar la siguiente igualdad en la población de personas:

$$\sigma^2_{X_p} = \sigma^2_T + \sigma^2_E.$$





Persona 1 (τ_1 = 64.30)

	Réplicas								
	1	2	3	4	5	6	7	8	⋯
τ_1	64.30	64.30	64.30	64.30	64.30	64.30	64.30	64.30	⋯
E_1	−2.31	2.59	−0.55	0.92	1.54	−1.57	3.65	0.36	⋯
X_1	61.99	66.89	63.75	65.22	65.84	62.73	67.95	64.66	⋯

Persona 2 (τ_2 = 81.87)

	Réplicas								
	1	2	3	4	5	6	7	8	⋯
τ_2	81.87	81.87	81.87	81.87	81.87	81.87	81.87	81.87	⋯
E_2	4.69	−1.71	1.63	−0.24	−3.40	3.69	−4.28	0.36	⋯
X_2	86.56	80.16	83.50	81.63	78.47	85.56	77.59	82.23	⋯

Población

Persona p	Puntuación observada *X*	Puntuación verdadera *T*	Error de medición *E*
1	61.99	64.30	−2.31
2	86.56	81.87	4.69
3	80.58	78.62	−1.95
4	68.43	69.39	0.96
5	59.11	58.07	−1.05
⋮	⋮	⋮	⋮
N	72.63	76.53	3.90

Figura 1. Representación gráfica del experimento mental que subyace a la teoría clásica de los tests.

Esta ecuación descompone la varianza observada en dos partes: varianza verdadera y varianza del error. En otras palabras, las diferencias que se observan entre las puntuaciones de las personas reflejan, por una parte, diferencias verdaderas entre las personas y, por otra parte, diferencias debidas a factores perturbantes.

Un concepto central en la TCT es la *confiabilidad*. Analizando la ecuación anterior, es claro que un instrumento es más confiable conforme las diferencias observadas son más diferencias verdaderas (y menos diferencias por errores de medición). De esta idea sigue la definición de la confiabilidad de un instrumento, representada por ρ, como la razón entre la varianza verdadera y la observada:

$$\rho = \frac{\sigma_T^2}{\sigma_X^{2\cdot}}$$

o, de forma equivalente:

$$\rho = 1 - \frac{\sigma_E^2}{\sigma_X^{2\cdot}}$$

Esta definición implica que la confiabilidad es un número entre 0 y 1 (siempre y cuando $\sigma_X^2 > 0$) y alcanza su máximo de 1 si $\sigma_X^2 = \sigma_T^2$ (toda la varianza observada es varianza verdadera) y su mínimo de 0 cuando $\sigma_X^2 = \sigma_E^2$ (todas las diferencias que se observan se deben a errores de medición).

Dado que la definición de la confiabilidad incluye un término desconocido (no se conocen las puntuaciones verdaderas ni su varianza), se desarrollaron métodos para *estimar* la confiabilidad a partir de una muestra. Los métodos más conocidos incluyen el método de formas paralelas, el test-retest, el método de dos mitades y el análisis interno (que incluye el famoso coeficiente α de Cronbach).[6] La exposición de estos métodos y su lógica se encuentra fuera del alcance de este artículo.

A la raíz cuadrada de la varianza de *E* se le llama *error estándar de medición*. Si se dispone de (una estimación de) la confiabilidad del instrumento y la varianza observada, se obtiene (una estimación de) el error estándar a través de la siguiente ecuación:

$$\sigma_E = \sigma_X \sqrt{1-\rho}. \qquad (3)$$

Teoría de respuesta al ítem: conceptos y modelos básicos

La TRI aproxima la medición de los constructos que un instrumento pretende evaluar de una forma radicalmente diferente que la teoría clásica. Mientras que la TCT considera la puntuación asociada con una prueba *en su globalidad* —nótese que el error de medición y la puntuación observada y verdadera se refieren a la prueba en su totalidad—, los modelos TRI analizan cómo las personas se comportan *en los elementos constituyentes* de la prueba, es decir, analizan las respuestas de cada persona en cada ítem de la prueba. Por lo tanto, la TRI es apropiada para analizar instrumentos compuestos de elementos más básicos (donde el ejemplo típico son los exámenes que consisten en diferentes preguntas) y menos como modelo para mediciones indivisibles, como la presión arterial o la temperatura corporal de una persona.

La TRI es una amplia familia de modelos psicométricos que comparten los siguientes supuestos básicos:

1. Subyacente a la prueba existen uno o más *constructos* o *rasgos latentes* (ciertas habilidades, actitudes, competencias, etc.) que intervienen cuando las personas responden a los ítems;
2. tanto las personas como los ítems tienen características relevantes (para los constructos mencionados) que se pueden resumir en uno o más *parámetros* (parámetros en la TRI son números que caracterizan un ítem o una persona);
3. las características de los ítems se definen independientemente de (es decir, existen sin referencia a) las personas, y viceversa, las características de las personas son independientes de los ítems;
4. es posible hacer una afirmación sobre la conducta de una persona específica en un ítem específico (por ejemplo, sobre la probabilidad de que lo acierte) tras la aplicación de una *regla* (generalmente, una función o una ecuación) que combina los parámetros de la persona y del ítem.

Los miembros de la familia TRI difieren entre sí respecto de (a) el número de rasgos latentes que suponen subyacentes a la prueba, (b) el número de parámetros que especifican para los ítems y, similarmente, el número de parámetros para las personas y (c) la regla que determina cómo combinar los parámetros de personas e ítems para llegar a una afirmación sobre la conducta observable en la prueba. Éste último tiene implicaciones directas para el formato de respuesta de los ítems (por ejemplo, ítems dicotómicos, con sólo dos respuestas posibles, *vs.* politómicos con múltiples categorías de respuesta) y el tipo de constructos subyacentes que el modelo permite analizar. A continuación, se introducen los conceptos básicos de la TRI a partir del modelo de Rasch.

El modelo de Rasch

Rasch[5] no fue el primero para proponer un modelo TRI: las ideas básicas de la TRI ya se formaron en los años 1940 y unos ocho años anteriores a Rasch, Lord había elaborado un modelo que se parece en varios sentidos al modelo de Rasch. Sin embargo, su elegancia, debido no sólo a la sencillez matemática y fácil aplicación, sino también a sus propiedades teóricas e implicaciones filosóficas,[11,15] hace que muchos expertos consideren el modelo de Rasch como el *primus inter pares* de la TRI.

Rasch modela la probabilidad de que una persona *p* (de alguna población de personas) conteste correctamente un ítem *i* (de alguna población de ítems). Por lo tanto, es para ítems con dos posibles respuestas, que típicamente se clasifican en "correcta" (o "acertar") e "incorrecta" (o "fallar"). A la respuesta de la persona *p* en el ítem *i* corresponde una variable aleatoria X_{pi}, que se define con los siguientes valores:

$$X_{pi} = \begin{cases} 1 \text{ si la persona } p \textit{ acierta} \text{ el ítem } i \\ 0 \text{ si la persona } p \textit{ falla} \text{ el ítem } i \end{cases}$$

Respecto de (a) el número de rasgos latentes, el modelo supone *unidimensionalidad*: un rasgo latente es

suficiente para describir el comportamiento de las personas en los ítems. Además, supone (b) que cada ítem y cada persona se caracterizan por sólo un parámetro: el parámetro del ítem *i* se representa por β_i, el parámetro de la persona *p* por θ_p, donde β_i y θ_p son números reales cualesquiera. Finalmente, en el modelo de Rasch, (c) la regla que combina β_i y θ_p para llegar a una afirmación sobre la probabilidad de que la persona acierte el ítem es

$$\Pr(X_{pi} = 1 \mid \theta_{pi}, \beta_i) = \frac{e^{\theta_p - \beta_i}}{1 + e^{\theta_p - \beta_i}} \quad (4a)$$

donde e es la base de los logaritmos naturales (e ≈ 2.718). Nótese que la expresión al lado derecho de la Ecuación (4a) transforma la diferencia $\theta_p - \beta_i$ (la cual, en principio, puede variar de $-\infty$ a $+\infty$) a un número entre 0 y 1, propio para una probabilidad. Dicha transformación se conoce como la transformación logística y entra, por ejemplo, también como función de enlace en modelos de regresión logística. Por lo tanto, el modelo de Rasch pertenece a la subfamilia de modelos logísticos dentro de la TRI.

Obviamente, puesto que el modelo considera únicamente dos categorías de respuesta, la probabilidad de que la persona *p* falle el ítem *i* es la complementaria de la Ecuación (4a):

$$\Pr(X_{pi} = 0 \mid \theta_p, \beta_i) = 1 - \Pr(X_{pi} = 1 \mid \theta_p, \beta_i),$$

lo cual, si se elabora algebraicamente, lleva a:

$$\Pr(X_{pi} = 0 \mid \theta_p, \beta_i) = \frac{1}{1 + e^{\theta_p - \beta_i}} \quad (4b)$$

Al considerar la probabilidad de acertar en función de la habilidad latente (es decir, al considerar en la Ecuación (4a) la *variable* θ en vez del *valor* θ_p que tiene la persona *p* en esta variable), se define la *curva característica del ítem*:

$$f_i(\theta) = \frac{e^{\theta - \beta_i}}{1 + e^{\theta - \beta_i}} \quad (5)$$

La curva característica define el modelo; o en otras palabras, se puede identificar un modelo TRI a partir de las curvas características de los ítems. En la gráfica izquierda de la **Figura** 2 se representan las curvas características de dos ítems en el modelo de Rasch. La única diferencia entre los ítems es su posición sobre la dimensión latente. Nótese que el parámetro de dificultad determina la posición de un ítem en el rasgo latente: β_i corresponde al nivel del rasgo para el cual la probabilidad de acertar el ítem *i* es 0.5. Efectivamente, de la Ecuación (5) sigue que, si $\theta = \beta_i$, entonces $f_i(\theta) = 0.5$.

Cabe resaltar algunas propiedades más de las curvas características en el modelo de Rasch. Primero, las curvas son crecientes —a mayor habilidad, mayor probabilidad de acertar—, lo cual es justo en un contexto donde el constructo subyacente es de rendimiento óptimo. Segundo, las asíntotas izquierda y derecha de la función son 0 y 1, respectivamente, lo cual quiere decir que la probabilidad de acertar un ítem se acerca a 1, conforme la habilidad de la persona incrementa y que, conforme la habilidad de la persona disminuye, la probabilidad de acertar cualquier ítem se acerca a 0. Tercero, como se observa en la gráfica derecha de la **Figura 2**, donde se representan las curvas características de varios ítems en un modelo Rasch, las curvas nunca intersectan. Se puede verificar en la Ecuación (5) que, si el ítem *i* es más fácil que el ítem *j* [$\beta_i < \beta_j$], entonces la probabilidad de acertar *i* es mayor que la probabilidad de acertar *j* [$f_i(\theta) > f_j(\theta)$], para *cualquier* nivel θ en el rasgo latente.

El supuesto de independencia local

Supongamos que se conocen el parámetro θ_p de una persona *p* y los parámetros β_i, β_j y β_k de tres ítems. En este caso, la Ecuación (4a) permite derivar la probabilidad, según el modelo de Rasch, de que la persona *p* dé la respuesta correcta en cada uno de los tres ítems. Sin embargo, la ecuación sólo proporciona dichas probabilidades *por separado*; no especifica cómo derivar la probabilidad *conjunta* de que, por ejemplo, la persona *p* acierte los ítems *i* y *j* y que falle el ítem *k*. Para llegar a afirmaciones sobre tales probabilidades, el modelo de Rasch (y la gran mayoría de otros modelos TRI) incluye un supuesto adicional: la *independencia local*. De este supuesto se desprende que la probabilidad conjunta equivale al producto de las probabilidades separadas.

Es esencial entender bien la cualidad de local en este supuesto; quiere decir que la independencia entre respuestas es condicional a la habilidad θ_p de la persona.

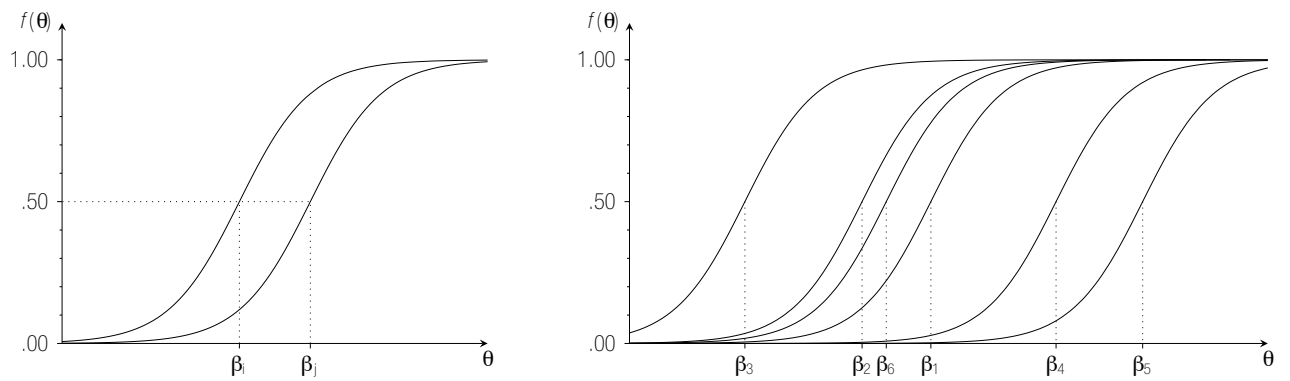


Figura 2. Ejemplos de curvas características en el modelo de Rasch.

Consideremos a dos personas *p* y *q* con el mismo nivel en el rasgo latente que responden los mismos tres ítems y supongamos que la persona *p* acertó los primeros dos ítems, mientras que la persona *q* los falló. A partir de esta información, ¿se concluirá que la persona *p* tendrá una probabilidad mayor que la persona *q* de acertar el tercer ítem? Si se acepta el supuesto de independencia local, la respuesta a esta pregunta es *no*. Bajo este supuesto, la probabilidad de acertar un ítem no cambiará a la luz de información adicional sobre las respuestas en otros ítems y depende únicamente del parámetro del ítem y el parámetro de la persona; puesto que en el ejemplo asumimos que $\theta_p = \theta_q$, la probabilidad de acertar el tercer ítem (y cualquier otro ítem) es la misma para ambas personas.

Si no se especificase que $\theta_p = \theta_q$, entonces la valoración de la probabilidad de acertar el tercer ítem sí sería diferente para las dos personas: después de haber observado que la persona *p* acertó los primeros dos ítems y la persona *q* los falló, es más plausible que $\theta_p > \theta_q$ y por lo tanto, es plausible que la persona *p* tenga una probabilidad mayor de acertar el tercer ítem. Lo importante en este razonamiento es que, en el modelo de Rasch y todos los demás modelos que incluyen independencia local entre sus supuestos, el ajuste en la probabilidad después de observar las respuestas en otros ítems se atribuye *exclusivamente* a la revaloración del nivel de la persona en el rasgo latente. Es decir, el nivel en el rasgo latente funciona como variable mediadora, lo cual implica que, si se mantiene θ_p fijo, entonces la probabilidad de acertar un ítem ya no se afecta por conocer las respuestas en otros ítems.

En resumen, el supuesto de independencia en los modelos TRI es local porque hace referencia a subgrupos de personas con idénticos valores en θ. Dentro de un grupo de personas en el cual todos tienen el mismo valor en θ, no hay correlación entre las variables X_i y X_j para cualquier par de ítems *i* y *j*; en otras palabras, si hay correlación, entonces es porque las personas difieren respecto de θ. La única causa de correlación entre los ítems es el rasgo latente. Esta consideración relaciona conceptualmente los supuestos de independencia local y de unidimensionalidad y ha llevado a algunos autores a la conclusión que el primero sigue directamente del segundo o que son empíricamente indistinguibles.[6,16,17]

Estimación de parámetros

Los valores de los parámetros de ítems y personas son inherentemente desconocidos. El objetivo principal de una aplicación del modelo de Rasch suele ser obtener estimaciones para estos parámetros a través de un análisis de las respuestas observadas en una muestra.

Existen varios métodos de estimación para modelos TRI. El método más común se conoce como *estimación por máxima verosimilitud* (en inglés: *maximum likelihood estimation*, MLE). MLE es una herramienta de estimación general en la estadística (introducida en los albores del siglo XX por R. A. Fisher)[18] y tiene una serie de propiedades teóricas atractivas que se sostienen en general bajo condiciones leves. Una exposición detallada de MLE en modelos TRI excede el alcance de este artículo; no obstante, se ilustra el principio con un ejemplo sencillo.

Supongamos que se conocen los parámetros $ß_1$, $ß_2$,…, $ß_6$ de los seis ítems que se graficaron en la parte derecha de la **Figura 2** y que se desea estimar el parámetro θ_p con base en las respuestas observadas de la persona *p* en estos ítems. La segunda y tercera columna de la **Tabla 1** muestran para cada ítem los valores del parámetro de dificultad y la respuesta que dio la persona *p*, respectivamente. Al aplicar MLE se considera la probabilidad de las respuestas observadas bajo los supuestos del modelo. En la cuarta columna de la tabla, se incluye la Fórmula (4a) o (4b) en función de la respuesta en cada ítem; por ejemplo, para el primer ítem, la cual no se contestó correctamente, la tabla aplica la Ecuación (4b). Nótese que en este ejemplo el único parámetro desconocido en las expresiones de la cuarta columna es θ_p. La idea fundamental de MLE es que se consideran diferentes valores para θ_p y, a continuación, se evalúa la probabilidad de las respuestas observadas. Esta última probabilidad es un indicador de la plausibilidad del parámetro y se llama la *verosimilitud* de θ_p.

Las últimas columnas de la **Tabla 1** muestran la verosimilitud para algunos valores ilustrativos de θ_p. Se observa, por ejemplo, que $\theta_p = 4$ es más verosímil que $\theta_p = 6$ para la respuesta observada en el primer ítem, mientras que para el segundo ítem la conclusión es al revés. Sin embargo, en vez de evaluar cada ítem por separado, se considera la verosimilitud del parámetro θ_p para el *patrón* de respuestas: bajo el supuesto de independencia local, la verosimilitud de θ_p dada las respuestas en los seis ítems, es el producto de las verosimilitudes de cada ítem. La última fila de la **Tabla 1** ilustra el cálculo para los cuatro valores de θ_p; cada resultado es el producto de las seis probabilidades precedentes en la misma columna.

Para examinar cómo la verosimilitud varía en función de todos los posibles valores de θ_p, se puede investigar la *función de verosimilitud*. La **Figura 3** muestra que la función de verosimilitud para el ejemplo anterior llega a su máximo si $\theta_p = 5.433$. Esto quiere decir que 5.433 es el más plausible entre todos los valores para θ_p. Se dice que $\hat{\theta}_p = 5.433$ es la estimación por máxima verosimilitud del parámetro θ_p (se pone un sombrero arriba del símbolo del parámetro para distinguir la estimación del valor verdadero).

En el contexto de estimar los parámetros de un modelo TRI, el problema resulta ser considerablemente más complejo debido a que se estiman simultáneamente múltiples parámetros. Sin embargo, la esencia del método sigue siendo la misma que lo expuesto para el caso anterior simple: se busca una solución para los parámetros desconocidos que tenga máxima verosimilitud a la luz de los datos observados. Cabe mencionar que se han desarrollado diversas variantes de MLE, principalmente para resolver algunos inconvenientes del método estándar. Por otro lado, en la última década ha incrementado sustancialmente el número de aplicaciones donde la estimación se realiza dentro del marco alternativo ofrecido por la estadística bayesiana.[19]

Para terminar esta sección, conviene señalar que el modelo de Rasch en su formulación general sufre de una falta de identificabilidad. Quiere decir que la solución de los parámetros no es única, que la función de verosimilitud no tiene uno, sino varios máximos. Un simple análisis

Tabla 1. Ilustración del principio de estimación por máxima verosimilitud para estimar el parámetro θ_p

Ítem	Parámetro del ítem	Respuesta observada	Prob. respuesta Ecuación (4)	Probabilidad de las respuestas observadas para algunos valores concretos de θ_p			
				$\theta_p = 4.0$	$\theta_p = 5.0$	$\theta_p = 5.5$	$\theta_p = 6.0$
1	$\beta_1 = 4.249$	$X_{p1} = 0$	$\frac{1}{1+e^{\theta_p-\beta_1}}$	.562	.321	.223	.148
2	$\beta_2 = 3.279$	$X_{p2} = 1$	$\frac{e^{\theta_p-\beta_2}}{1+e^{\theta_p-\beta_2}}$	.673	.848	.902	.938
3	$\beta_3 = 1.627$	$X_{p3} = 1$	$\frac{e^{\theta_p-\beta_3}}{1+e^{\theta_p-\beta_3}}$	.915	.967	.980	.988
4	$\beta_4 = 6.014$	$X_{p4} = 1$	$\frac{e^{\theta_p-\beta_4}}{1+e^{\theta_p-\beta_4}}$	.118	.266	.374	.497
5	$\beta_5 = 7.235$	$X_{p5} = 0$	$\frac{1}{1+e^{\theta_p-\beta_5}}$	.962	.903	.850	.775
6	$\beta_6 = 3.620$	$X_{p6} = 1$	$\frac{e^{\theta_p-\beta_6}}{1+e^{\theta_p-\beta_6}}$	.594	.799	.868	.915
			Verosimilitud $\ell(\theta_p)$:	.023	.051	.054	.048

de la Ecuación (4) provee el argumento: la probabilidad de acertar o fallar un ítem depende de los parámetros de la persona y del ítem únicamente a través de su diferencia $\theta_p - \beta_i$. Por lo tanto, cuando se dispone de una estimación de los parámetros (θ_1, θ_2,..., θ_N, β_1, β_2,..., β_n) a partir de las respuestas de *N* personas en *n* ítems, se puede construir otra solución sumando una constante *c* arbitraria a los parámetros de todas las personas y todos los ítems, como sigue:

$$\theta_p^* \leftarrow \theta_p + c$$
$$\beta_i^* \leftarrow \beta_i + c$$

En este caso, la solución (θ_1^*, θ_2^*,...,θ_N^*,β_1^*,β_2^*,...,β_n^*) produce las mismas probabilidades a través de la Ecuación (4) que la solución original, puesto que $\theta_p^* - \beta_i^* = \theta_p - \beta_i$ para cualquier combinación de una persona *p* y un ítem *i*. Una forma común para resolver esta indeterminación del modelo consiste en añadir la restricción que la media aritmética de los parámetros β_i de los ítems sea 0.

Función de información

El método de MLE que se introdujo en la sección anterior proporciona una estimación puntual de los parámetros. En muchas ocasiones, es deseable tener también una indicación de la precisión de la estimación, por ejemplo, en términos de un intervalo de confianza para el valor verdadero del parámetro. Un teorema en la teoría de MLE muy relevante al respecto dice que el valor de un parámetro estimado por máxima verosimilitud se puede considerar aproximadamente como una extracción de una distribución normal cuya media es el valor verdadero del parámetro y cuya varianza es el inverso de la función de información.[20]

La función de información se define en términos del valor esperado de la segunda derivada del logaritmo de la función de verosimilitud. Aplicado al parámetro θ en el modelo de Rasch, se puede derivar que la información proporcionada por un test de *n* ítems para estimar θ se da por:

$$I(\theta) = \sum_{i=1}^{n} f_i(\theta)[1-f_i(\theta)] \quad (6)$$

donde $f_i(\theta)$ se define como en la Ecuación (5). El producto en la suma del lado derecho, $f_i(\theta)$ [1 − $f_i(\theta)$], se llama la función de información del ítem *i* sobre el parámetro θ (generalmente presentada por $I_i(\theta)$. El resultado en la Ecuación (6) implica que la información proporcionada por el test en su totalidad es una suma simple de las informaciones proporcionadas por los ítems. Esto se muestra en la **Figura** 4, donde se representa la función de información para los seis ítems y del test en su totalidad para el ejemplo de la **Tabla** 1. Además, la figura ilustra que la función de información de cada ítem es máxima cuando θ coincide con el parámetro de dificultad. Esto quiere decir que un ítem proporciona más información para estimar la habilidad de las personas cuyo parámetro se encuentra cerca del parámetro de dificultad del ítem y menos para las personas que se encuentran lejos del ítem en la dimensión latente. Extendiendo esta idea al test en su totalidad, se concluye que proporciona más

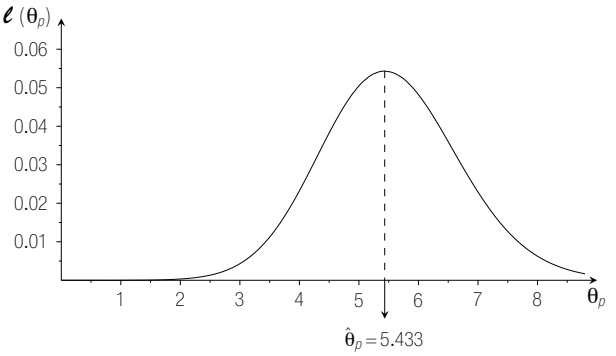


Figura 3. Función de verosimilitud para el parámetro θ_p del ejemplo en la **Tabla 1**.

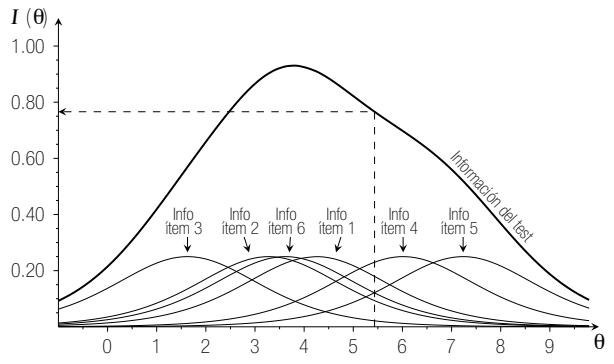


Figura 4. Función de información de los seis ítems de la **Tabla 1**, por separado y conjuntamente (del test en su totalidad) para estimar el parámetro θ. La línea discontinua muestra la información del test ($I(\theta)$ = 0.766) para un valor de θ = 5.433.

información para estimar los valores de θ que se encuentran entre las β_i de los ítems.

El teorema mencionado al inicio de esta sección permite derivar intervalos de confianza para el parámetro de interés. Para calcular el intervalo de confianza asociado con $\hat{\theta}_p$ = 5.433 en el ejemplo anterior, se considera primero el error estándar de la estimación:

$$\sigma_{\hat{\theta}_p} \approx \frac{1}{\sqrt{I(\theta)}} \qquad (7)$$

Puesto que se desconoce el valor verdadero θ_p de la persona *p* en el ejemplo, se utiliza el valor estimado $\hat{\theta}$= 5.433 para calcular la información. Como se puede leer en la **Figura 4**, I(5.433) = 0.766, por lo cual

$$\sigma_{\hat{\theta}_p} \approx \frac{1}{\sqrt{0.766}} = 1.143$$

A continuación, se utiliza el método común para derivar un intervalo de confianza para la media de una distribución normal (con desviación estándar conocida); se obtiene el siguiente intervalo de 95% para el valor verdadero del parámetro θ_p en nuestro ejemplo:

$$\begin{aligned} &[\hat{\theta}_p - 1.96 \times \sigma_{\hat{\theta}_p}, \hat{\theta}_p + 1.96 \times \sigma_{\hat{\theta}_p}] \\ \approx &[5.433 - 1.96 \times 1.143, 5.433 + 1.96 \times 1.143] \\ = &[3.193 , 7.673] \end{aligned}$$

Nótese que este intervalo de confianza es muy amplio, lo cual se debe a que el ejemplo conformaba sólo seis ítems.

En aplicaciones reales, donde se desconocen los parámetros de los ítems (contrario al caso de nuestro ejemplo), se reemplazan también los $\hat{\beta}_i$ en la Ecuación (6) por las respectivas estimaciones . Cuando se requieren intervalos de confianza para los parámetros β_i, se puede aplicar un procedimiento similar a la que se acaba de presentar para los θ_p (aunque la función de información $I(\beta)$ es otra). Por otro lado, si se ha optado por una estimación dentro del marco bayesiano, se examina la distribución posterior del parámetro de interés (y específicamente su varianza) para evaluar la precisión de las estimaciones.

Dos modelos alternativos: el 2PL y 3PL

En esta sección se describen brevemente otros dos modelos TRI, que relajan los supuestos del modelo de Rasch en el sentido que permiten que los ítems difieran en otra(s) característica(s) que sólo el parámetro de dificultad. En otras palabras, en dichos modelos cada ítem se cuantifica en dos o tres parámetros, lo cual explica sus nombres: 2PL (modelo logístico de 2 parámetros) y 3PL (modelo logístico de 3 parámetros).[21] En muchos otros aspectos, como el supuesto de unidimensionalidad e independencia local, el 2PL y 3PL son similares al modelo de Rasch.

El 2PL añade un parámetro de *discriminación* a cada ítem, el cual se representa por α_i. En el panel superior izquierdo de la **Figura 5**, se representan las curvas características de dos ítems, *i* y *j*, que tienen la misma dificultad ($\beta_i = \beta_j$) pero que difieren en su parámetro de discriminación: $\alpha_j > \alpha_i$. Se observa que la curva del ítem *j* es más pronunciada cerca de su posición en el rasgo latente. En particular, comparando cualquier par de personas *p* y *q*, con parámetros $\theta_p < \beta_i = \beta_j < \theta_q$ (es decir, uno se encuentra por debajo de los parámetros de dificultad, el otro por encima), se cumple la siguiente desigualdad:

$$\Pr(X_{qj} = 1) - \Pr(X_{pj} = 1) > \Pr(X_{qi} = 1) - \Pr(X_{pi} = 1).$$

Esta expresión significa que, si los parámetros de dificultad de dos ítems coinciden, entonces la diferencia entre ambas personas respecto de su probabilidad de acertar los ítems es más grande en el ítem con mayor grado de discriminación. En la gráfica izquierda, la diferencia entre las probabilidades de acertar de las personas *p* y *q* es .769 − .289 = .480 en el ítem *j*, pero sólo .630 − .401 = .229 en el ítem *i*. Efectivamente, el ítem *j* discrimina más entre (las probabilidades de acertar de) estas dos personas. De forma alternativa, la probabilidad de encontrar una diferencia entre las respuestas de ambas personas (que uno acierte y el otro falle) es más grande en el ítem *j* que en el ítem *i*.

La ecuación matemática de la curva característica de un ítem *i* en el 2PL es la siguiente:

$$f_i(\theta) = \frac{e^{\alpha_i(\theta-\beta_i)}}{1 + e^{\alpha_i(\theta-\beta_i)}}$$

Para que $f_i(\theta)$ sea creciente, se añade la restricción que $\alpha_i > 0$. Alineado con la interpretación anterior, la ecuación enseña que el parámetro de discriminación encoge (si $\alpha_i < 1$) o estira (si $\alpha_i > 1$) la diferencia entre los parámetros de la persona y del ítem. En el panel superior derecho de la **Figura 5**, se grafican las curvas características de una familia de ítems que difieren tanto en dificultad como en discriminación.

El tercer parámetro para los ítems, que se introduce en el modelo 3PL, se suele denominar el parámetro de *seudo-adivinación*. Dicho parámetro, que se representa por γ_i, cambia la asíntota izquierda de la curva característica: mientras que en el modelo de Rasch y el 2PL, las personas con una habilidad muy baja tienen (casi) nula probabilidad de acertar el ítem, en el 3PL está probabilidad se acerca a γ_i, donde γ_i satisface la restricción: $0 \le \gamma_i \le 1$. La curva característica de un ítem *i* según el 3PL se da por:

$$f_i(\theta) = \gamma_i + (1 - \gamma_i)\frac{e^{\alpha_i(\theta-\beta_i)}}{1 + e^{\alpha_i(\theta-\beta_i)}} \qquad (8)$$

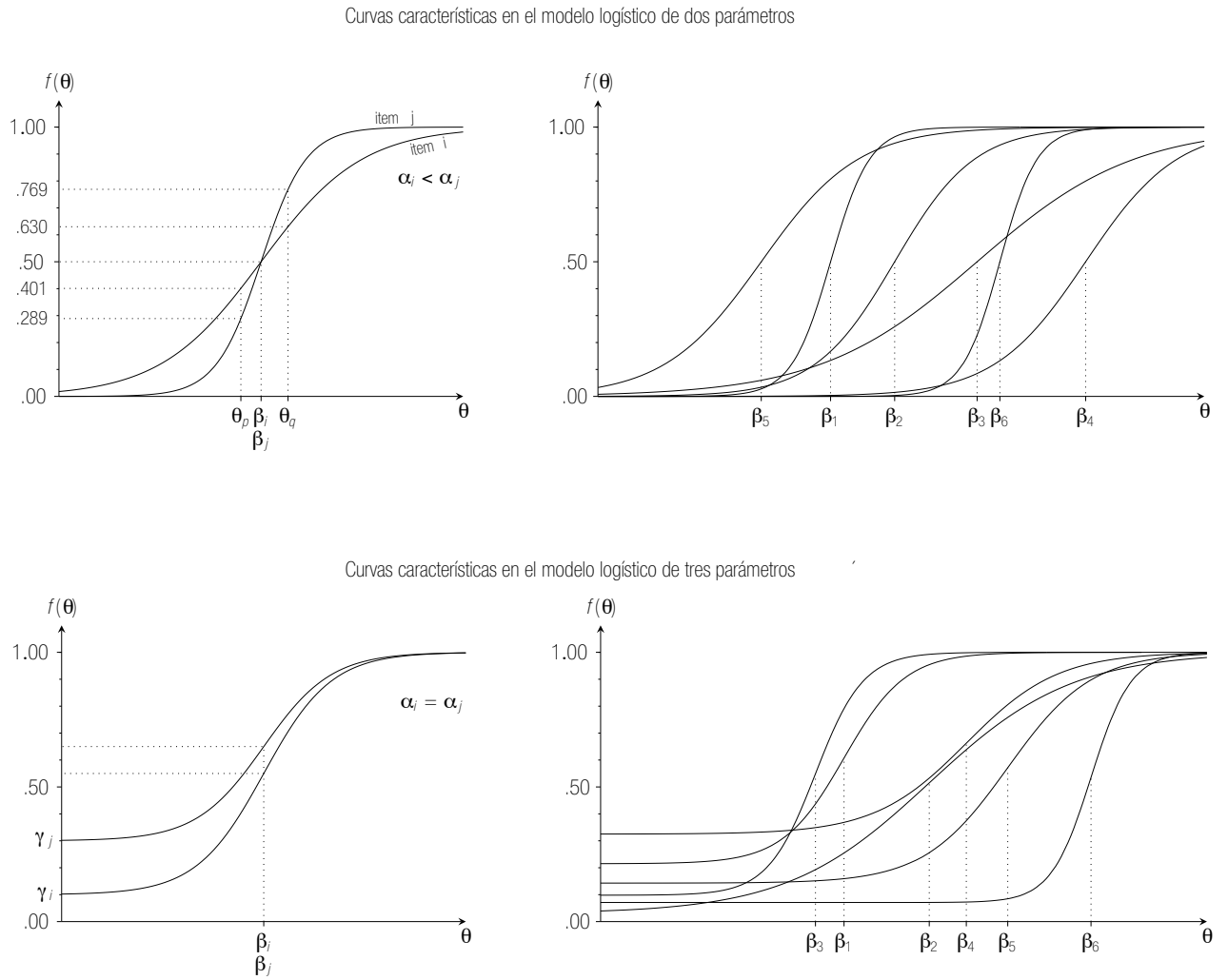


Figura 5. Ejemplos de curvas características en los modelos logísticos de dos y tres parámetros.

En la parte inferior de la **Figura** 5 se ejemplifican unas curvas características típicas del 3PL. El panel izquierdo ilustra el efecto del parámetro de seudo-adivinación comparando dos ítems cuyos valores en los otros dos parámetros son iguales. Nótese que β_i en el 3PL ya no corresponde con la habilidad para la cual la probabilidad de acertar es .50 (sino con la posición donde esta probabilidad es .50 + γ_i / 2).

Se suele interpretar γ_i como la probabilidad de acertar el ítem *i* en caso de que se "desconozca la respuesta", por lo cual este modelo parece ser adecuado para el análisis de ítems de opción múltiple, o bien, ítems que se pueden acertar adivinando, sin conocer la respuesta. Sin embargo, la interpretación que se adhiere a un parámetro, por ejemplo, interpretar γ_i en términos de "la probabilidad de adivinar correctamente" no es parte (de la definición formal) del modelo. Es posible que las respuestas en un ítem *i* se describan bien por la Ecuación (8), aunque las personas *no* adivinen, sino que, por ejemplo, hayan llegado a la respuesta correcta con base en un razonamiento erróneo.[22]

En general, la interpretación de los parámetros en el 2PL y 3PL es menos unívoca que en el modelo de Rasch. Por ejemplo, se mencionó anteriormente que Rasch implica que si $\beta_i < \beta_j$, entonces el ítem *i* es más fácil que el ítem *j* para *todas* las personas. En el 2PL, sin embargo, esta interpretación no es necesariamente correcta, como se ilustra con los ítems 3 y 6 en la gráfica en el panel superior derecho de la **Figura** 5: a pesar de que el ítem 6 tiene el parámetro de dificultad más *grande* que el ítem 3, la probabilidad de acertarlo es más *alta* (es decir, el ítem 6 es más *fácil*) para una parte significativa de la dimensión latente (en específico, para las personas con niveles altos en θ). En el 3PL, la interpretación es aún más confusa. Es posible que el ítem *i* sea más *fácil* que el ítem *j* en términos de su parámetro de dificultad ($\beta_i < \beta_j$), sin embargo, que el ítem *i* sea más *difícil* en el sentido que su curva característica se encuentre por debajo de —o bien, la probabilidad de acertarlo sea más baja que— la del ítem *j* para *todos* los niveles de la habilidad subyacente (para un ejemplo, compárense los ítems 2 y 4 en el panel inferior derecho). Este tipo de

consideraciones han sido objeto de un debate intenso a favor y en contra del 3PL.

Al terminar esta introducción de los modelos TRI básicos, cabe señalar que el modelo de Rasch es un caso especial del modelo 2PL, que se obtiene restringiendo los parámetros de discriminación a 1. Similarmente, el 2PL es un caso especial que se obtiene por la restricción de $\gamma_i = 0$ para cada ítem *i*. En la siguiente sección se considera la bondad de ajuste de un modelo a los datos. La jerarquía entre los tres modelos introducidos en esta sección implica que el modelo 3PL es más flexible y generalmente tiene mejor ajuste a los datos, mientras que el modelo de Rasch es el más exigente y fácilmente se rechaza. Como se explicó en el párrafo anterior, esta flexibilidad viene con el precio de una interpretación menos clara de los parámetros.

Análisis teórico de las diferencias entre ambos enfoques psicométricos

Desde un punto de vista teórico, las ventajas de la TRI son difíciles de negar. Un análisis TRI generalmente proporciona información más detallada, más sofisticada y con un sustento teórico más sólido. En esta sección se discuten brevemente cuatro ventajas de la TRI sobre la TCT.

Interpretación de las puntuaciones

En la gran mayoría de los casos, el objetivo final de una medición es hacer inferencias sobre alguna habilidad abstracta o algún constructo subyacente a la prueba utilizada. Una pregunta que la TCT (en su formulación original) deja sin contestar es hasta qué grado la puntuación en la prueba contiene información sobre este constructo subyacente. Incluso en el caso de que se midiese con exactitud la puntuación verdadera, no es claro qué conclusiones se pueden sacar sobre un constructo latente con base en el resultado obtenido; el modelo de la TCT simplemente no especifica la relación entre la habilidad latente que supuestamente se mide y el resultado observado en el test. Los parámetros en los modelos TRI, por otro lado, tienen una relación directa con la dimensión que se pretende medir, lo cual conlleva a una interpretación más clara de los resultados. Por ejemplo, θ *coincide con* (una cuantificación de) la habilidad abstracta subyacente a la prueba.

La falta de la especificación de la relación entre la puntuación en el test y la habilidad que se pretende medir tiene, además de su relevancia teórica, varias implicaciones prácticas. En primera instancia, no es evidente cuál es el nivel de medición de la puntuación (observada o verdadera) de la TCT. Si, por ejemplo, la puntuación de la persona *p* en una prueba es mayor que la de la persona *q*, entonces ¿se puede concluir que *p* tiene más de la habilidad que mide el test que la persona *q*? En la TCT, es muy común calcular la puntuación en el test por la suma de puntuaciones en los ítems, la cual en el caso de ítems binarios corresponde con el número de respuestas correctas. Se puede derivar que, si los supuestos del modelo Rasch son ciertos, entonces la respuesta a la pregunta anterior es afirmativa: la puntuación obtenida por el número de respuestas correctas refleja el orden entre las personas en la dimensión subyacente que se mide. Sin embargo, si los ítems difieren en discriminación (como en el 2PL), es posible que una persona con una puntuación más alta que otra persona reciba una estimación más baja para su parámetro de habilidad. Con el mismo razonamiento, se puede cuestionar incluso si las puntuaciones de la TCT cumplen con los requisitos más básicos de medición.

Una segunda implicación apunta al significado de las puntuaciones en la TCT con relación a algún estándar o criterio de decisión. Los exámenes en México se aprueban usualmente al obtener el 60% de la calificación máxima, pero el enfoque tradicional de la TCT carece de ilación sobre lo que "sabe" la persona que logra aprobar el examen con esta calificación. Los modelos de la TRI, al contrario, ponen las habilidades de las personas en la misma dimensión que las dificultades de los ítems y permiten concluir cuáles son los ítems que una persona domina (donde "dominar" tiene un significado preciso; en el modelo Rasch, por ejemplo, se dice que una persona *p* domina un ítem *i* si $\theta_p > \beta_i$ y entonces la probabilidad de que lo acierte es mayor que .50). Es decir, a partir de un análisis TRI, se obtiene información sobre el nivel de la persona en el constructo subyacente en relación con los ítems incluidos en el test.

Chequeo y falsabilidad del modelo

Algunos autores han defendido la TCT refiriéndose a los leves supuestos del modelo, que "no requieren evaluaciones estrictas del ajuste a los datos".[23] Sin embargo, es cuestionable abogar a favor del uso de un modelo (o una teoría científica en general) con el argumento de que hay pocas posibilidades que los datos lo rechacen. La filosofía de la ciencia (y especialmente el principio popperiano de falsabilidad) adopta una posición opuesta.[24]

Más fundamental es la objeción de que la teoría clásica (y específicamente el supuesto central de que ε_p se extrae aleatoriamente de una distribución de probabilidad y que su efecto en τ_p es aditivo) no es comprobable. Además, se ha reconocido que otros supuestos (menos esenciales) son poco realistas y/o se violan en la práctica (como el supuesto en la Ecuación 2 o el supuesto de que *E* es independiente de *T* en la Ecuación 1).[25,26] Sin embargo, entre los usuarios de la TCT existe la cultura de no preocuparse por los supuestos del modelo y proceder como si fuesen correctos.

Los supuestos en la TRI, por otro lado, generalmente son más exigentes y aunque se admite que ningún modelo es una representación perfecta del proceso cognitivo que subyace a los datos, se considera esencial evaluar, a través de pruebas estadísticas de bondad de ajuste, si es justificable mantener el modelo como hipótesis para los mismos. Comúnmente, ajustar un modelo TRI implica un proceso iterativo: a partir de un modelo inicial, (a) se evalúa el ajuste global a los datos y en caso de que resulte inaceptable, (b) se aplican pruebas tendientes a hallar violaciones específicas, con base en las cuales (c) se realizan modificaciones precisas; después se regresa al punto (a) hasta obtener un modelo final con un ajuste satisfactorio. En la siguiente sección se ejemplificará cómo se pueden adaptar los modelos TRI para acomodar violaciones comunes en el contexto de la evaluación educativa en las ciencias de la salud.

Error estándar de medición

Como se discutió anteriormente (véase la Ecuación 2), la TCT añade a sus supuestos que el error estándar de medición es una constante para cualquier nivel de la puntuación verdadera. La TRI, por otro lado, permite que el error estándar varíe en función de la habilidad subyacente (véase la Ecuación 7) y, efectivamente, evidencia que la precisión asociada con una medición *no* es constante en toda la escala, sino menor a los extremos. Además, intuitivamente parece lógico que una medición sea menos confiable en caso de una discordancia entre la dificultad global de la prueba y el nivel de la persona cuyo nivel se desea medir (es decir, si la prueba es demasiado fácil o difícil).

Si el supuesto de un error estándar de medición parejo no es correcto, la estimación de σ_E (a través de, por ejemplo, una estimación de la confiabilidad por el coeficiente α de Cronbach y una aplicación de la Fórmula 3) corresponde aproximadamente con la media de los errores estándares individuales.[25] Por lo tanto, para unas personas el error estándar global es una subestimación de su error estándar individual; para otras es una sobreestimación. En consideración de que la precisión de la medición en el contexto de la evaluación educativa no es igualmente importante para todos los niveles en la escala —subestimar una calificación verdadera de 60% con 2% generalmente tiene implicaciones mucho más graves que cometer un error del mismo tamaño cuando la calificación verdadera sea 82%—, conviene concentrar las fuerzas para que el error estándar se minimice alrededor de la(s) línea(s) divisoria(s) relacionada(s) con las decisiones que se planean tomar con base en el instrumento.

Invarianza de los parámetros

La diferencia más saliente entre los dos enfoques principales de la psicometría probablemente es la invarianza de los parámetros en la TRI. Quiere decir que, si los supuestos de un modelo TRI se cumplen para una población de personas y una población de ítems, entonces:

a. las propiedades de los ítems (es decir, sus parámetros como dificultad y discriminación) o de una prueba en su totalidad (por ejemplo, la función de información) no cambian al considerarlos o aplicarlos en diferentes muestras de personas; las propiedades serían las mismas en una muestra de personas dotadas y una muestra de personas menos capaces. En la TCT esto no es el caso: los índices asociados con una prueba generalmente son distintos en diferentes muestras de personas. Por ejemplo, la confiabilidad suele ser más baja en un grupo de personas más homogéneo y el grado de dificultad de un ítem (el cual se define, en el caso de que la respuesta se codifique de forma binaria, como la proporción de personas que lo acierta) es diferente en grupos de personas capaces y menos capaces.
b. los parámetros de las personas son los mismos independientemente de la muestra de ítems que se incluyan en la prueba; no importa, por ejemplo, que la prueba tenga mayoritariamente ítems fáciles, o bien, difíciles, las θs de las personas son idénticas en cualquier caso. Las características de las personas en la TCT —más notablemente, sus puntuaciones verdaderas— *no* son idénticas en diferentes pruebas: si la versión A de un examen incluye más preguntas fáciles que otra versión B, las puntuaciones verdaderas de la versión A serán más altas.

Es importante insistir en la interpretación correcta de la propiedad de invarianza de los parámetros en modelos TRI, ya que a veces publicaciones sobre el tema difunden una interpretación errónea. La invarianza o la independencia de la muestra *no* implica que la *estimación* de los parámetros de los ítems sea independiente de la muestra de personas. Esto sólo es cierto para un subgrupo (importante) de modelos TRI, a saber la familia de modelos tipo Rasch, la cual además del modelo de Rasch introducido anteriormente, incluye varios modelos que comparten las propiedades especiales del modelo de Rasch (véase el libro de Wright y Stone[27] o el editado por Fischer y Molenaar[28] para una discusión más profunda). Además, incluso para modelos tipo Rasch, aunque el valor esperado de la estimación de los parámetros es independiente de la muestra, la *precisión* de la estimación no lo es, como se mostró en la sección donde se introdujo la función de información.

Desde la perspectiva clásica es, en principio, imposible separar la influencia de la versión utilizada para una prueba, por un lado, y la contribución de las personas, por otro lado, en las calificaciones obtenidas. Gracias a la invarianza de los parámetros de los ítems, la TRI permite comparar el rendimiento de distintos (grupos de) individuos aunque contestaron diferentes versiones de una prueba. Un tipo de aplicaciones que explota máximamente esta propiedad son los *tests adaptativos informatizados* (TAI).[29] Lo típico de un TAI, para lo cual se requieren una computadora equipada de un software especial y un banco amplio de ítems calibrados, es que se estima el nivel θ de la persona después de *cada* respuesta y que se elige entre los ítems restantes del banco el más adecuado (generalmente el más informativo condicional a la estimación actual de θ) para presentar como el siguiente. Las pruebas populares *Test of English as Foreign Language* (TOEFL) y *Test of English for International Communication* (TOEIC) tienen una versión adaptativa.[30]

Los supuestos de modelos TRI *vs*. la realidad de la evaluación educativa en medicina

En esta sección se reconsideran los supuestos de los modelos TRI básicos y se contrastan con las circunstancias en las que se realizan las evaluaciones típicas en medicina. Ampliando la perspectiva y considerando las posibles soluciones cuando un modelo muestra un ajuste deficiente a los datos, se pueden clasificar generalmente las soluciones en dos grupos, dependiendo de si se sitúa la causa del problema en el modelo, o bien, en los datos. En el primer caso, la estrategia para remediar el mal ajuste consiste en modificar el modelo; en el segundo caso, se procede a la resolución del problema cambiando los datos, más específicamente, eliminando ítems y/o personas con índices de ajuste problemáticos. Ciertamente, un análisis

psicométrico puede revelar problemas específicos en algunos ítems, que después de un escrutinio más a fondo, puede llevar a la conclusión de que no son aptos, por ejemplo, porque traen una interpretación ambigua. Asimismo, se puede justificar la eliminación de una persona después de examinar a fondo sus respuestas al instrumento y constatar, por ejemplo, que no respondió con seriedad a la tarea.

Sin embargo, demasiadas veces se consigue un ajuste aceptable al modelo contemplado después de la eliminación de un porcentaje significativo de los ítems, con la justificación que tienen un "mal ajuste" o "valores extremos/inaceptables para los parámetros". En general, es importante tratar los datos "con respeto" y utilizar con mucha cautela la estrategia de dejar fuera del análisis parte de los datos para obtener un ajuste a un modelo estadístico. Comúnmente, a cada ítem en la prueba se le concedió una importancia en la fase de la construcción de la prueba, por lo cual la eliminación de ítems generalmente afecta la validez de contenido. Aunque en casos muy particulares las propiedades únicas de un modelo psicométrico (como el modelo de Rasch) pueden justificar su preponderancia sobre los datos, en la mayoría de las aplicaciones el modelo es secundario a los datos. Es decir, si se encuentra un ajuste deficiente para una parte significativa de las personas e ítems, casi siempre es más apropiado reconsiderar el modelo en vez de censurar los datos.

En el resto de esta sección se discuten brevemente algunos modelos TRI alternativos que se han desarrollado precisamente para responder a las implicaciones que tienen ciertos contextos de evaluación educativa para los supuestos de la TRI.

Adivinar aleatoriamente en ítems de opción múltiple

En la práctica de la evaluación educativa, el 3PL es el modelo estándar para analizar exámenes de opción múltiple desde la perspectiva de la TRI. Sin embargo, varios autores[31,32] han criticado el supuesto inherente de que la probabilidad de acertar un ítem adivinando (lo cual es la interpretación común del parámetro γ_i) únicamente depende del ítem y es constante para todas las personas. Han argumentado que, en caso de que no se sepa la respuesta, también las características de la persona -como su nivel de habilidad- afectan qué tan atractivas le parecen las distintas opciones de respuesta. Por ejemplo, para una persona con un nivel *muy* bajo en la habilidad subyacente, las alternativas pueden parecer igualmente atractivas, así que la probabilidad de acertar el ítem adivinando se acerca a $1/k$ (con k el número de opciones de respuesta); por otro lado, una persona de un nivel más alto, aunque no sepa la respuesta correcta, a lo mejor puede identificar uno o más distractores (por lo cual se incrementaría la probabilidad de acertar adivinando), mientras que también es posible que una alternativa incorrecta engañe a personas de cierto nivel de habilidad (en cuyo caso, la probabilidad de acertar el ítem se disminuiría). En resumen, es poco plausible suponer que la opción correcta atraiga igualmente a todas las personas que no saben la respuesta. San Martín y cols.[32] presentaron evidencia que este supuesto del 3PL se viola más fácil en preguntas que permiten que el conocimiento se manifieste de forma gradual.

Bock[33] y Thissen y Steinberg[34] propusieron alternativas para el análisis de ítems de opción múltiple. Al asignar parámetros a las distintas opciones de respuesta, estos modelos definen una curva característica para cada *opción*. La **Figura 6** muestra para un ítem de cuatro alternativas cómo varía la probabilidad de elegir cada opción en función de la habilidad θ. Observando la curva característica de la opción correcta, se nota que inicialmente la probabilidad de acertar el ítem *disminuye* conforme θ incrementa, hasta cierto punto desde el cual la probabilidad de acertar se acerca a 1 para θ yendo a infinito. Esta gráfica ejemplifica cómo las características de los distractores conllevan a una curva característica fundamentalmente distinta a la del 3PL.

Además de que representan mejor el proceso cognitivo de responder a un ítem de opción múltiple, los modelos mencionados anteriormente acarrean otra ventaja: aprovechan toda la información presente en los datos. El 3PL no diferencia las respuestas en los distractores de un mismo ítem (analiza la información dicotomizada de respuesta correcta *vs.* incorrecta). Estudios han mostrado que incluir esta información conlleva a una mejor estimación de la habilidad de las personas.[35,36]

Unidimensionalidad

Todos los modelos examinados hasta ahora suponen que sólo un rasgo latente subyace a las respuestas observadas: para fines de la prueba estudiada, cada persona *p* se puede reducir a una posición θ_p en la dimensión subyacente, la cual se combina con los parámetros de los ítems para conocer la probabilidad de acertarlos. Sin embargo, en pocas ocasiones la evaluación educativa de estudiantes en medicina (y en otras áreas) es unidimensional; generalmente, un examen sondea diversas subáreas en las cuales las diferencias entre estudiantes se manifiestan de forma distinta. Esto no es sólo el caso en exámenes que explícitamente incluyen preguntas de diversas áreas clínicas y ciencias básicas,[37] sino también en exámenes de una asignatura en particular y aun cuando cubren sólo parte de la materia (como en los exámenes parciales).

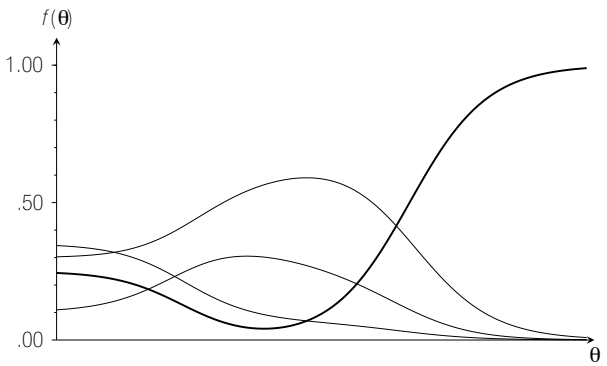


Figura 6. Curvas características para un ítem de cuatro opciones en el modelo para ítems de opción múltiple de Thissen y Steinberg.[34] La curva gruesa representa la opción correcta.

Para tomar en cuenta la multidimensionalidad de los exámenes en el contexto educativo, se puede adoptar el marco de la *TRI multidimensional* (véase el libro de Reckase[38] para una visión general). Un modelo TRI multidimensional reconoce que las respuestas observadas resultan de la interacción de un número (fijo) de constructos (dimensiones). Típicamente incluye (a) parámetros específicos de cada dimensión (tanto para personas como para ítems), y (b) una regla que combina (los parámetros asociados con) las distintas dimensiones para llegar a afirmaciones probabilísticas sobre el comportamiento de las personas en los ítems. Actualmente, la TRI multidimensional es un área muy activa de investigación y las revistas especializadas publican a menudo artículos sobre este tipo de modelos.

Por el número de parámetros involucrados en un modelo multidimensional, frecuentemente resulta cómodo añadir restricciones específicas para dar forma a la multidimensionalidad. Considérese por ejemplo, el modelo propuesto por Gibbons y Hedeker,[39] el cual es relativamente sencillo y fácil de aplicar y, además, puede tener mucha relevancia para la evaluación educativa. El modelo requiere que los ítems se clasifiquen previamente en *m* grupos que corresponden con *m* diferentes áreas. Al nivel de la prueba total, se supone que (a) un constructo *general* interviene en la respuesta a cualquier ítem, independientemente del grupo al que pertenece, y (b) que con cada grupo de ítems se asocia un constructo *específico*, que únicamente afecta las respuestas a los ítems de este grupo. Nótese que en este modelo los ítems sólo tienen parámetros para dos constructos (el general y uno de los específicos), mientras que las personas ocupan una posición en cada una de las *m* + 1 dimensiones. Aunque el modelo original de Gibbons y Hedeker sólo contiene dos niveles de jerarquía (general *vs.* específico), es relativamente sencillo extenderlo para incorporar estructuras jerárquicas más complejas (que permiten acomodar, por ejemplo, áreas y subáreas).[40] La ventaja del modelo propuesto por estos autores es que la estimación de los parámetros es factible incluso con un número grande de áreas y subáreas.

Independencia local

Como se explicó anteriormente, el supuesto de independencia local significa que considerando fijo el valor θ_p, se valora igual la probabilidad de que la persona *p* acierte el ítem *i*, independientemente de si haya o no acertado el ítem *j*. En la literatura se han descrito situaciones particulares que conllevan una violación del principio de independencia local. La situación que probablemente tiene más relevancia en el contexto de la educación médica se presenta cuando un examen incluye una serie de casos clínicos y para cada uno de estos casos clínicos se realizan dos o más preguntas.

Muy ilustrativo al respecto es el siguiente caso clínico (del área de Urgencias médicas) que se presentó en un examen sumativo de altas consecuencias en nuestro medio. Después de haber introducido el cuadro clínico, se realizó una primera pregunta para la cual los sustentantes eligiesen el diagnóstico más probable entre:

A. Gastroenteritis probablemente infecciosa.
B. Apendicitis aguda.
C. Absceso hepático amebiano.

En la segunda pregunta, se les propusieron los siguientes manejos terapéuticos:

A. Cloranfenicol parenteral.
B. Apendicectomía.
C. Metronidazol parenteral.

de los cuales tuvieron que escoger el más apropiado. Ambas preguntas tenían la opción B como la correcta y el análisis mostró que eran relativamente fáciles. Considérese ahora un participante con un alto nivel de habilidad y su probabilidad de contestar correctamente la segunda pregunta. Puesto que tiene un valor alto en θ y que la pregunta es fácil (es decir, con un grado de dificultad ß bajo), se valorará alta la probabilidad de que acierte. Si ahora se dispone de la información adicional que esta persona ha llegado al diagnóstico equivocado de gastroenteritis, ¿se valoraría igualmente alta la probabilidad de acertar la segunda pregunta? Es obvio que ningún médico decida a una apendicectomía después de haber llegado a un diagnóstico de gastroenteritis. En otras palabras, la respuesta en la primera pregunta condiciona la respuesta a la segunda. Este caso constituye un ejemplo muy claro de la violación del principio de independencia local; analizar este examen (que incluyó más casos clínicos del mismo tipo) con un modelo como el 3PL lleva a un ajuste deficiente y a estimaciones erróneas de los niveles de θ.[41]

Afortunadamente, se han desarrollado alternativas dentro de la TRI para tomar en cuenta este tipo de dependencias entre ítems. La mayoría de estas soluciones requiere que se identifique *a priori* (posiblemente después de un análisis preliminar) los ítems entre los cuales puede existir una dependencia. El método más apropiado para incluir estas dependencias en el modelo se elige básicamente en función de la estructura general de la prueba y el número de ítems que son interdependientes. Si por ejemplo, la prueba consiste en una mezcla de preguntas aisladas y varios casos con dos o tres preguntas, puede ser adecuada la extensión propuesta por Hoskens y De Boeck.[42] Cuando, por otro lado, la prueba consiste en un número de *testlets* (es decir, clusters de ítems interdependientes) y el número de ítems en cada testlet es relativamente grande, puede ser más indicada la metodología de Wainer y cols.[43,44]

Discusión y conclusiones

En este artículo se revisaron los conceptos fundamentales de la TCT y de la TRI, los cuales constituyen actualmente los dos paradigmas principales de la psicometría. Se compararon los méritos de ambos enfoques y además, se examinaron los modelos de la TRI y la plausibilidad de sus supuestos a la luz de las prácticas típicas en el área de la evaluación educativa en las ciencias de la salud. A partir de este análisis se llega a la conclusión de que los supuestos de los modelos TRI tradicionales por muchas razones no son compatibles con dichas prácticas y que el entusiasmo limitado para la TRI (y el consiguiente predominio de la TCT) en el campo de la evaluación educativa en medicina parcialmente se debe a las dificultades que acompañan el ajuste de un modelo TRI a datos recopilados en el contexto educativo.

Respecto de la TCT, cabe mencionar que este artículo se limitó a la formulación original del modelo clásico y que no indagó en las contribuciones más recientes que han propuesto soluciones para remediar algunos de los problemas del modelo original. Aunque estas soluciones generalmente son parches *ad hoc* que no resuelven el problema fundamental (por ejemplo, las fórmulas para derivar múltiples errores estándares de medición para la misma población o los métodos para lograr una equiparación de las puntuaciones obtenidas en diferentes versiones de una prueba), dos extensiones aportaron una perspectiva distinta al enfoque tradicional: la perspectiva del análisis factorial de los ítems en una prueba[13] y la teoría de la generalizabilidad.[45,46] La primera se acerca mucho a la TRI debido a que realiza un análisis a nivel de los ítems y examina la estructura de una prueba apelando a uno o más factores subyacentes (que son conceptualmente idénticos a los rasgos latentes en la TRI). La teoría de la generalizabilidad, por otro lado, extiende el modelo clásico examinando las posibles fuentes de variación y sus contribuciones relativas a los datos observados. En este sentido, explicita cómo pueden variar las condiciones en el experimento mental que se introdujo en la sección sobre la teoría clásica.

Esperamos que este artículo haya aclarado que la TRI es una familia muy extensa de modelos, a pesar de que tocó nada más una selección muy limitada de los mismos. Por ejemplo, todos los modelos revisados (excepto los modelos que analizan las *k* categorías de respuesta en las preguntas de opción múltiple) son para datos binarios; existen extensiones de modelos para preguntas cuyas respuestas se codifican en múltiples categorías ordenadas e incluso para el caso de calificaciones continuas. Otros modelos que no se tocaron en este artículo incluyen los que permiten el análisis de datos provenientes de múltiples jueces (como por ejemplo en el examen oral largo ante paciente real, que es una de las alternativas para aprobar el examen profesional en la Facultad de Medicina), modelos para analizar datos longitudinales para investigar cambios en el rasgo latente durante el tiempo, y modelos multidimensionales que permiten investigar las reglas de decisión y los esquemas implícitos utilizados por los médicos (por ejemplo, los criterios necesarios y suficientes que llevan a los diagnósticos en ciertos contextos). El alcance de modelos TRI para el análisis de datos de evaluación educativo es virtualmente ilimitado.

Este artículo no prestó mucha atención a algunas limitaciones prácticas de la TRI. Entre las más importantes se encuentran las muestras más amplias que generalmente se requieren para obtener estimaciones estables para los parámetros; aunque el tamaño requerido por un modelo TRI generalmente depende de su complejidad y el número de parámetros incluidos —modelos con más parámetros generalmente requieren muestras más grandes—, casi siempre la TRI resulta más exigente que la TCT al respecto. También la complejidad matemática y cuestiones técnicas (como el uso de programas especializados para la estimación) pueden disuadir a los investigadores que no son expertos en psicometría y poner trabas a la toma de decisiones adecuadas en aplicaciones concretas.[23]

Concluyendo, cabe reconocer que se enfatizaron más las diferencias que las similitudes entre los dos enfoques psicométricos. Efectivamente, varios autores han buscado tender un puente entre los dos paradigmas.[13,47,48] Especialmente, la reconsideración del modelo clásico dentro de la perspectiva del análisis factorial llevó a cierta reconciliación entre la TCT y la TRI y mostró, para modelos particulares, que ambos enfoques pueden llegar a resultados y conclusiones similares en aplicaciones prácticas y que las diferencias se relacionan más con perspectivas filosóficas distintas.

Agradecimientos

El autor agradece a Florina Gatica Lara por sus sugerencias sobre una versión anterior y a Alma Jurado Núñez por compartir sus ideas y reflexiones que enriquecieron el texto actual.

Financiamiento

Ninguno.

Conflicto de intereses

El autor declara no tener ningún conflicto de intereses.

Presentaciones previas

Ninguna.

Referencias

1. Spearman C. Demonstration of formulae for true measurement of correlation. Am J Psychol 1907;18:161–169.
2. Spearman C. Correlation calculated from faulty data. Br J Psychol 1910;3:271–295.
3. Guttman L. A basis for scaling qualitative data. Am Social Rev 1944;9:139–150.
4. Lord F. A theory of test scores. Psychometric Monograph 7. Richmond, VA: Psychometric Corporation; 1952.
5. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press; 1980.
6. Abad FJ, Olea J, Ponsoda V, García C. Medición en ciencias sociales y de la salud. Madrid, España: Síntesis; 2011.
7. Muñiz J. Introducción a la teoría de respuesta a los ítems. Madrid, España: Pirámide; 1997.
8. Muñiz J. Teoría clásica de los tests. 2a ed. Madrid, España: Pirámide; 2002.
9. Crocker L, Algina J. Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston; 1986.
10. de Ayala RJ. The theory and practice of item response theory. New York, NY: Guilford; 2009.
11. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
12. Furr RM, Bacharach VR. Psychometrics: An introduction. Thousand Oaks, CA: Sage; 2008.
13. McDonald RP. Test theory: A unified treatment. Mahwah, NJ: Erlbaum; 1999.
14. Novick MR. The axioms and principal results of classical test theory. J Math Psychol 1966;3:1–18.
15. Andrich D. Controversy and the Rasch model: A characteristic of incompatible paradigms? Med Care 2004;42(S1):7–16.
16. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.
17. Ip EH. Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. Br J Math Stat Psychol 2010;63:395–416.

18. Aldrich J. R. A. Fisher and the making of maximum likelihood 1912 – 1922. Stat Sci 1997;12:162–176.
19. Fox JP. Bayesian item response modeling: Theory and applications. Nueva York, NY: Springer; 2010.
20. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics. 3a ed. Nueva York, NY: McGraw-Hill; 1974.
21. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. En: Lord FM, Novick MR, editores. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley; 1968. p. 396–479.
22. Glas CAW, Verhelst ND. Een overzicht van itemresponsmodellen. En: Eggen TJHM, Sanders PF, editores. Psychometrie in de Praktijk. Arnhem, Holanda: Cito; 1993. p. 179–238.
23. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educ Meas Issues Pract 1993;12(3):38–47.
24. Echeverría J. Filosofía de la ciencia. 2a ed. Madrid, España: Akal; 1998.
25. Raju NS, Price LR, Oshima TC, et al. Standardized conditional SEM: A case for conditional reliability. Appl Psychol Meas 2007;31:169–180.
26. Zimmerman DW, Williams RH. Chance success due to guessing and non-independence of true scores and error scores in multiple-choice tests: Computer trials with prepared distributions. Psychol Rep 1965;17:159–165.
27. Wright BD, Stone MH. Best test design: Rasch measurement. Chicago, IL: MESA; 1979.
28. Fischer GH, Molenaar IW, editores. Rasch models: Foundations, recent developments, and applications. Nueva York, NY: Springer-Verlag; 1995.
29. Olea J, Ponsoda V. Tests adaptativos informatizados. Madrid, España: UNED; 2013.
30. Olea J, Abad FJ, Ponsoda V, et al. eCAT-Listening: Design and psychometric properties of a computerized adaptive test on English Listening. Psicothema 2011;23:802–807.
31. Samejima F. A new family of models for the multiple choice item. Reporte de Investigación 79-4. Knoxville, TN: Universidad de Tennessee, Departamento de Psicología; 1979.
32. San Martín E, del Pino G, De Boeck P. IRT models for ability-based guessing. Appl Psychol Meas 2006;30:183–203.
33. Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 1972;37:29–51.
34. Thissen D, Steinberg L. A response model for multiple choice items. Psychometrika 1984;49:501–519.
35. Levine MV, Drasgow F. The relation between incorrect option choice and estimated ability. Educ Psychol Meas 1983;43:675–685.
36. Thissen DM. Information in wrong responses to the Raven Progressive Matrices. J Educ Meas 1976;13:201–214.
37. Delgado-Maldonado L, Sánchez-Mendiola M. Análisis del examen profesional de la Facultad de Medicina de la UNAM: Una experiencia de evaluación objetiva del aprendizaje con la teoría de respuesta al ítem. Inv Educ Med 2012;1:130–139.
38. Reckase MD. Multidimensional item response theory. Nueva York, NY: Springer; 2009.
39. Gibbons RD, Hedeker DR. Full-information item bi-factor analysis. Psychometrika 1992;57:423–436.
40. Rijmen F. Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. En: von Davier M, Hastedt D, editores. Issues and Methodologies in Large-Scale Assessments. vol. 4 of IERI Monograph Series. Hamburgo, Alemania: 2011. p. 59–74.
41. Chen CT, Wang WC. Effects of ignoring item interaction on item parameter estimation and detection of interacting items. Appl Psychol Meas 2007;31:388–411.
42. Hoskens M, De Boeck P. A parametric model for local dependence among test items. Psychol Methods 1997;2:261–277.
43. Wainer H, Kiely GL. Item clusters and computerized adaptive testing: A case for testlets. J Educ Meas 1987;24:185–201.
44. Wainer H, Wang X. Using a new statistical model for testlets to score TOEFL. J Educ Meas 2000;37:203–220.
45. Cronbach LJ, Gleser GC, Nanda H, et al. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. Nueva York, NY: Wiley; 1972.
46. Brennan RL. Generalizability theory. Nueva York, NY: Springer; 2010.
47. Lord FM. Applications of item response theory to practical testing problems. Mahwah, NJ: Lawrence Erlbaum; 1980.
48. Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika 1987;52:393–408.