Capítulo 17

UNA INTRODUCCIÓN A LA TEORÍA DE RESPUESTA AL ÍTEM PARA EL ANÁLISIS PSICOMÉTRICO DE EXÁMENES

José J. Naveja, Iwin Leenen

"Todos los modelos están equivocados, pero algunos son útiles."
George Box

INTRODUCCIÓN

En el capítulo anterior se ha abordado el enfoque de la teoría clásica de los tests (TCT) para el análisis psicométrico de los resultados de un examen. La teoría clásica está comprendida por un modelo único cuyo objetivo es estimar la "calificación verdadera" de cada uno de los sustentantes en el examen. Otro enfoque en la psicometría, más reciente, es la teoría de respuesta al ítem (TRI; en ciertas publicaciones, teoría de respuesta al reactivo). Contrario a la TCT, donde el interés está en la calificación global del examen, la TRI analiza las respuestas de los sustentantes en cada uno de los reactivos del examen, suponiendo explícitamente que estas permiten hacer una inferencia sobre una o más características psicológicas de los individuos. En la literatura psicométrica, se suele utilizar el término "rasgos latentes" para dichas características psicológicas que no son directamente observables. (En publicaciones más antiguas se hablaba de la "teoría del rasgo latente", en vez de la TRI.) Entonces, en la TRI, en contraste con la TCT, el foco se localiza en el (los) rasgo(s) latente(s) de interés y no en el instrumento que se utiliza para medirlo(s). Otra diferencia con la teoría clásica, es que la TRI incluye una gama muy amplia de modelos que permite acomodar diferentes tipos de reactivos en el examen (de respuesta abierta, de opción múltiple, etc.), así como probar distintos supuestos sobre el proceso psicológico subyacente que relaciona la respuesta en el reactivo con el (los) rasgo(s) latente(s) que se desea(n) medir.

En el presente capítulo, abordamos en primera instancia una breve reseña histórica de las ideas que dieron lugar a la TRI. Posteriormente, introducimos los principios teóricos que fundamentan el marco conceptual de esta teoría. Después, presentamos algunos ejes que permiten organizar la familia de los modelos TRI y profundizamos en algunos miembros de esta familia (sin duda, los más conocidos y utilizados para el análisis psicométrico de

exámenes). La sección subsecuente aborda ciertos temas relevantes en la aplicación práctica de los modelos TRI: estimación, bondad de ajuste y confiabilidad. Concluimos el capítulo con una discusión de las ventajas e inconvenientes del enfoque de la TRI.

ANTECEDENTES HISTÓRICOS DE LA TEORÍA DE RESPUESTA AL ÍTEM

Wim van der Linden (2016) destaca el trabajo de Alfred Binet como una contribución pionera a la TRI. A principios del siglo XX, el gobierno francés le confirió a Binet el encargo de desarrollar una prueba estandarizada para distinguir a niños con discapacidad intelectual de niños con inteligencia normal pero escasa motivación. El resultado de este proyecto fue el primer test de inteligencia, publicado por Binet en 1905 en colaboración con Théodore Simon. La innovadora metodología que Binet propuso estaba muy adelantada a su época en diferentes aspectos. En primera instancia, reconoció que la variable de interés, la inteligencia, no es directamente observable, sino latente. A Binet le interesaba desarrollar una escala para presentar sus resultados, pero concluyó que no habría una escala natural para medir un rasgo latente como la inteligencia; entonces, aplicó sus ítems a participantes de diferentes grupos etarios y asignó a cada ítem el valor de la edad cronológica con el cual el 75% de los examinados podían responderlo correctamente. Este sistema permitía estimar la edad mental de cada participante según su desempeño. Así, Binet utilizó la misma escala para medir tanto a los ítems del test como al desempeño de los participantes, lo cual es una característica fundamental de la TRI.

Dos décadas después del planteamiento de Binet, Louis Thurstone consideró necesario medir la inteligencia con su propia escala, que sería abstracta e inobservable por definición. Esto evitaría ambigüedades en la interpretación de la escala en la comparativa entre la edad mental y la cronológica: aún si la edad mental se mantuviera estable en un periodo de tiempo, el cociente intelectual disminuiría automáticamente con solo aumentar la edad cronológica. La separación de la escala del rasgo latente de otras características observables permitió a Thurstone extender los conceptos desarrollados en el contexto de la medición de la inteligencia a la medición de otras características psicológicas intangibles, como las actitudes o los rasgos de personalidad. No obstante, en el trabajo de Thurstone, y de otros investigadores que retomaron sus ideas, todavía se pueden apreciar ambigüedades en la conceptualización del rasgo latente y el uso de los modelos para su estimación.

Un paso importante en la evolución hacia la TRI consistió en utilizar funciones matemáticas para modelar la probabilidad de observar cierta respuesta en un ítem en función de la característica latente de interés en las personas. Los primeros que adoptaron estas "funciones probabilísticas de respuesta" fueron Frederic Lord (1952), con su "modelo de la ojiva normal", y George Rasch (1960), con el modelo logístico. Los avances de Rasch fueron más profundos: también propuso métodos para la estimación de parámetros y la evaluación de la bondad de ajuste del modelo. Hoy en día, el modelo de Rasch es todavía uno de los modelos más prominentes de la TRI.

Muchos otros psicómetras posteriores a Lord y Rasch hicieron avances subsecuentes que permitieron diversificar la gama disponible de modelos, por ejemplo, extendiendo los

modelos para ítems con múltiples opciones de respuesta, o incorporando más parámetros de interés, tales como la adivinación de respuestas. En la siguiente sección entraremos en los principios teóricos que comparten todos los modelos de la TRI.

PRINCIPIOS TEÓRICOS DE LOS MODELOS DE LA TEORÍA DE RESPUESTA AL ÍTEM

Todos los modelos psicométricos que se han desarrollado en el marco de la TRI comparten una serie de propiedades o principios que justamente los definen como miembros de la familia TRI. La primera propiedad es que todos estos modelos se enfocan en la probabilidad de que una persona *p*, al momento de responder un reactivo *i*, dé una respuesta en la categoría *j* del reactivo; dicha probabilidad se escribe $\Pr(Y_{pi} = j)$. Lo anterior quiere decir que un modelo TRI permite derivar, con base en sus supuestos, $\Pr(Y_{pi} = j)$. (Se utiliza el índice *p* para referirse de manera general a cualquier persona de la población para la cual se construyó la prueba; asimismo, los índices *i* y *j* se refieren de manera general a cualquier reactivo de la prueba y a cualquier categoría de respuesta en este reactivo, respectivamente. Un modelo TRI categoriza de cierta manera a las posibles respuestas; véase el inicio de la siguiente sección.)

La segunda propiedad que todos los modelos TRI comparten es la especificación de uno o más parámetros para cada persona y, de manera separada, uno o más parámetros para cada reactivo. Un parámetro es una característica teórica, ya sea de la persona o del reactivo, que se expresa a través de un valor numérico. La interpretación psicológica del parámetro depende del modelo concreto; en los modelos que asignan solo un parámetro a cada persona, este parámetro indica el nivel de la persona en el rasgo latente que se quiere medir con la prueba. En un examen de inglés, por ejemplo, el parámetro asignado a la persona *p* es un número que indica el dominio de inglés de esta persona. En la siguiente sección explicaremos para algunos modelos concretos de la TRI qué significado tienen los parámetros.

Finalmente, un modelo TRI siempre especificará una "regla" de cómo se combinan el o los parámetros de la persona *p* y el o los parámetros del reactivo *i* para llegar a la probabilidad $\Pr(Y_{pi} = j)$, comúnmente a través de una función matemática. Lo anterior se puede escribir como:

$$\Pr(Y_{pi} = j) = f(\text{parámetro (s) de la persona } p, \text{ parámetro(s) del ítem } i) \qquad (1)$$

En los modelos descritos en la siguiente sección, se especificará cada vez esta ecuación matemática, la cual constituye el núcleo del modelo. A continuación, se introducen algunos conceptos y términos clave de los modelos TRI: unidimensionalidad, función característica e independencia local.

Unidimensionalidad

No todos los modelos TRI son unidimensionales; sin embargo, los primeros modelos propuestos –y los más conocidos– incluyen el supuesto de unidimensionalidad. Esto quiere decir que se supone que solo hay un rasgo latente que explica las diferencias (sistemáticas) entre las respuestas de distintas personas en los reactivos de la prueba. Por ejemplo, en un examen de matemáticas, unidimensionalidad implica que solo la habilidad matemática explica, de manera sistemática, el por qué ciertos sustentantes dan mejores respuestas en el examen que otros. Aun cuando otras habilidades, como la comprensión lectora, pueden también ser relevantes o necesarias para resolver el examen (por ejemplo, para entender las instrucciones y las preguntas), un modelo unidimensional supondría que estas otras características no causan diferencias entre (las respuestas que dan) distintos sustentantes. Como mencionamos anteriormente, los modelos unidimensionales asignan solo un parámetro a cada persona (su nivel en el único rasgo latente que la prueba mide); el valor numérico de este parámetro posiciona a la persona en una "dimensión" (es decir, una línea) donde posiciones más altas corresponden con niveles más altos en el rasgo latente.

Nótese que, para muchos exámenes en la universidad, es poco plausible el supuesto de unidimensionalidad. Considérese, por ejemplo, el examen de ingreso a la UNAM; esta prueba evalúa múltiples áreas de conocimiento (español, matemáticas, historia, literatura, etcétera) y, en este sentido, es de esperar que las diferencias entre los sustentantes reveladas por el examen se deban a múltiples rasgos latentes. Incluso para un examen dedicado a un área específica, como las matemáticas, es posible que subyazcan múltiples rasgos latentes (por ejemplo, habilidad en álgebra, geometría, probabilidad, etcétera.). Si estos múltiples rasgos latentes causan diferencias entre personas y, además, son relativamente independientes (es decir, si las personas pueden tener un nivel de dominio relativamente alto en un rasgo, pero relativamente bajo en otro), entonces el supuesto de unidimensionalidad se incumple. En este caso, a veces se utilizan modelos unidimensionales como aproximación, aunque es más apropiado hacer uso de modelos multidimensionales de la TRI. Para los lectores interesados, se recomienda el libro de Reckase (2010).

Función Característica

Como se mencionó anteriormente, un modelo TRI permite calcular, a través de la ecuación básica (la cual se escribió de manera genérica en la Ecuación 1), la probabilidad de cierta respuesta con base en los parámetros de la persona y del ítem. Cuando se considera esta probabilidad como una función del (de los) rasgo(s) latente(s) implicado(s) por la prueba, se obtiene la *función característica* para esta categoría de respuesta en el ítem. En otras palabras, la función característica describe, para una categoría de respuesta en un reactivo en particular, cómo la probabilidad cambia si el nivel de la persona en el (los) rasgo(s) latente(s) cambia(n). En la frase anterior, el nivel de la persona se considera una variable (es decir, ya no hace referencia a una persona en particular) y la función característica, entonces, considera *en general* cambios en el (los) rasgo(s) latente(s) y cómo estos cambios afectan la probabilidad de una respuesta particular en un ítem específico. En la siguiente sección, donde se presentan varios modelos TRI concretos, se darán ejemplos de las funciones características.

Concluimos esta sección sobre la función característica con dos comentarios. Primero, en los modelos unidimensionales, donde solo un rasgo latente subyace a las respuestas observadas, es común representar la función característica en una gráfica, como la que se encuentra en el panel (a) de la Figura 1. En esta gráfica la abscisa (eje horizontal) indica el nivel en el rasgo latente y la ordenada (eje vertical) la probabilidad de responder en la categoría del ítem bajo consideración; el trazo describe cómo la probabilidad cambia en función del nivel en el rasgo latente. En casi todos los modelos TRI este trazo es una curva y, por eso, es común hablar de la "curva característica" como sinónimo de la "función característica".

Segundo, en los modelos para ítems dicotómicos (es decir, modelos que consideran solo dos posibles categorías de respuesta en cada ítem, por ejemplo, respuesta correcta *versus* respuesta incorrecta; véase la siguiente sección), se suele hablar de "la función característica del ítem", sin hacer referencia a una categoría de respuesta en particular. Sin embargo, la referencia es implícita y comúnmente alude a la categoría de respuesta "correcta" (más en general, la categoría que está asociada con niveles superiores en el (los) rasgo(s) latente(s). De esta manera, la función característica del ítem describe cómo *la probabilidad de acertar* el ítem cambia en función del nivel en el (los) rasgo(s) latente(s). Debe ser claro que la probabilidad de responder en la otra categoría del ítem (es decir, fallarlo) es el complemento de la probabilidad de acertar: si la probabilidad de acertar el ítem (para cierto nivel de la persona) es 0.70, entonces la probabilidad de fallarlo es 1 – 0.70 = 0.30.

Independencia Local

La ecuación básica de un modelo TRI permite calcular, para cualquier persona, su probabilidad de cierta (categoría de) respuesta para cada ítem. Por ejemplo, para el caso de exámenes de opción múltiple, hay modelos que permiten calcular la probabilidad de que la persona elija en el primer reactivo la opción B (por ejemplo, 30%), la probabilidad de que elija D en el segundo (por ejemplo, 80%), la probabilidad de que elija A en el tercero (por ejemplo, 50%), etc. Sin embargo, además de estas probabilidades para ciertas respuestas en reactivos *individuales*, los modelos TRI también especifican la probabilidad de cierto *patrón de respuestas*. En el ejemplo anterior del examen de opción, podríamos estar interesados en la probabilidad del patrón (B,D,A); es decir, la probabilidad de esta *combinación* particular de respuestas (en los primeros tres reactivos).

Para esta probabilidad, se incorpora de cierta forma un supuesto de independencia en el modelo. La forma más común consiste en suponer que una vez que se conoce el valor en el (los) parámetro(s) de la persona (es decir, su nivel en el (los) rasgo(s) latente(s) bajo estudio), conocer su respuesta en uno de los reactivos no dará información adicional acerca de qué pudo haber contestado en otro ítem. Supongamos en el ejemplo anterior que el examen de opción múltiple evalúa, como único rasgo, "dominio de inglés"; si *con base en el nivel de la persona en este rasgo* el modelo indica que la probabilidad de elegir la opción correcta en el último reactivo es igual a 50%, entonces esta probabilidad no cambiaría si conociéramos las respuestas en los reactivos previos; incluso si la persona acertara todos los reactivos previos (o los fallara todos), la probabilidad de acertar el último reactivo sigue siendo 50%.

El supuesto se llama *independencia local* (o *independencia condicional*), precisamente porque las respuestas en los otros reactivos no importan *una vez que se haya tomado en cuenta el nivel de la persona en el rasgo latente*. Con las probabilidades mencionadas en el primer párrafo de esta sección, el supuesto de independencia local implicaría que la probabilidad del patrón de respuestas (B,D,A) se obtiene multiplicando las probabilidades de los ítems individuales ($0.30 \times 0.80 \times 0.50 = 0.12$).

LA FAMILIA DE LOS MODELOS TRI

Se han desarrollado decenas (si no centenas) de modelos en el marco de la TRI. Varios autores (por ejemplo, Thissen y Steinberg, 1986) propusieron taxonomías para dar orden y estructura en esta jungla de modelos TRI. La mayoría de estas taxonomías se basan en las propiedades teóricas de los modelos. A continuación, presentamos unos ejes organizadores que sirven para diferenciar entre los distintos modelos y que se basan en criterios que son más relevantes para su aplicación:

- *Número de categorías de respuesta: modelos para ítems dicotómicos versus modelos para ítems politómicos.*
  Los primeros modelos TRI consideraban únicamente ítems dicotómicos: ítems con solo dos categorías de respuesta. Como se solían aplicar en el área de las habilidades cognitivas, estas categorías se denominaban "respuesta correcta" *versus* "respuesta incorrecta".[1] Cabe mencionar que estos modelos también se pueden utilizar cuando originalmente hay múltiples categorías de respuesta (como en los reactivos de opción múltiple de tres o más opciones), pero en estos casos se deben dicotomizar las respuestas antes de llevar a cabo el análisis, es decir, juntar ciertas opciones en la misma categoría (por ejemplo, todos los distractores de un ítem de opción múltiple se juntan en la categoría "respuesta incorrecta").

  Desarrollos posteriores permitieron analizar ítems politómicos, que en general tienen m categorías de respuesta (donde $m \geq 2$). Cuando se trata de ítems politómicos, adicionalmente se diferencia entre el caso donde las categorías estén ordenadas *a priori* (por ejemplo, cuando la calificación en los reactivos sea 0, 1, 2 o 3, y calificaciones más altas representen una mejor respuesta) *versus* el caso donde no exista un orden *a priori* (como entre las opciones de respuesta en un reactivo de opción múltiple; aunque la respuesta correcta se considera la mejor, los distractores entre sí suelen no estar ordenados). En general, analizar ítems que tienen múltiples (más de 2) respuestas con modelos para

[1] Sin embargo, esta nomenclatura no quiere decir que la aplicación de estos modelos esté limitada para el análisis de exámenes. También se pueden aplicar para analizar encuestas que miden actitudes, donde las categorías de respuesta podrían corresponder con "de acuerdo" y "en desacuerdo", o cuando se miden rasgos de personalidad, donde los reactivos son descriptores de personalidad y las categorías de respuesta son "me describe bien" o "no me describe bien".

ítems politómicos lleva a resultados más confiables en comparación con la aproximación de dicotomizar las respuestas y analizarlas con modelos para ítems dicotómicos, debido a que la dicotomización implica una pérdida de información.

- *Dimensionalidad del rasgo latente: modelos unidimensionales versus modelos multidimensionales.*
  En la sección anterior se explicó el supuesto de unidimensionalidad, que es parte de muchos de los modelos TRI. Para aplicaciones prácticas es importante considerar la plausibilidad de que solo un rasgo latente subyace a las respuestas de una prueba. En general, esta consideración guía la decisión sobre el uso de modelos unidimensionales *versus* modelos multidimensionales.

- *Forma de la función característica del ítem: modelos de dominancia versus modelos de proximidad.*
  Este capítulo se enfoca en el uso de la TRI para el análisis de exámenes. En este caso, el (los) rasgo(s) latente(s) de interés suele(n) hacer referencia a cierta(s) habilidad(es) y generalmente un nivel más alto en esta(s) habilidad(es) conlleva una probabilidad más alta de tener un mayor puntaje en el reactivo. Esto significa que la función característica (para la categoría más alta) del ítem es una función creciente. Modelos con funciones características crecientes se llaman *modelos de dominancia* (para tener un mayor puntaje en el reactivo se requiere una mayor dominancia del (los) rasgo(s) latente(s)). Efectivamente, para el análisis de exámenes se suelen utilizar modelos de dominancia.

  Sin embargo, es importante tener en mente que para cierto tipo de reactivos y cierto tipo de rasgos latentes puede ser inapropiada una función característica creciente. Consideremos, como ejemplo, el siguiente reactivo en un cuestionario para conocer la opinión de los estudiantes acerca de las medidas que se deben aplicar a profesores que acosan a las estudiantes:

  > *La sanción adecuada para profesores que acosen sexualmente a sus estudiantes es quitarles parte de su sueldo por seis meses.*

  Suponiendo que este reactivo se responde con "de acuerdo" o "en desacuerdo" y que el rasgo latente que subyace las respuestas es actitud (de rechazo) leve *versus* actitud severa hacia el acoso, es de esperar que tanto las personas con una actitud muy severa como aquellas con una actitud muy leve tengan una probabilidad baja de responder "de acuerdo" (en el primer caso porque la sanción mencionada les puede parecer insuficiente, en el segundo porque pueden considerar la sanción demasiado dura); y que personas con una actitud intermedia tendrían la probabilidad más alta de estar de acuerdo. En otras palabras, para este tipo de ítems puede ser más apropiada una función característica unimodal, es decir, una función que indica que la probabilidad aumenta hasta cierto punto y después baja. Los *modelos de proximidad* o *punto ideal* tienen funciones características

de este tipo. El lector interesado en estos modelos puede consultar los capítulos en la Sección VI en el libro de van der Linden (2016).

En el resto de esta sección, presentamos algunos de los modelos TRI más conocidos y más relevantes para el análisis de exámenes. Todos son modelos unidimensionales de dominancia. Los primeros tres son modelos para reactivos dicotómicos, con puntajes 0 (respuesta incorrecta) y 1 (respuesta correcta); el último es para reactivos politómicos con categorías ordenadas (por ejemplo, preguntas abiertas donde el puntaje mínimo es 0 y el puntaje máximo es 2 o más).

El Modelo de Rasch (o el Modelo Logístico de Un Parámetro)

Este modelo, propuesto por Rasch (1960), es uno de los más simples de los que se utilizan en la TRI. Asigna a cada persona p un parámetro, representado por θ_p, y también a cada ítem un parámetro, β_i, y combina estos parámetros (ambos números reales) a través de la siguiente ecuación para hacer una afirmación probabilística de que la persona acierte el ítem:

$$\Pr(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}. \tag{2a}$$

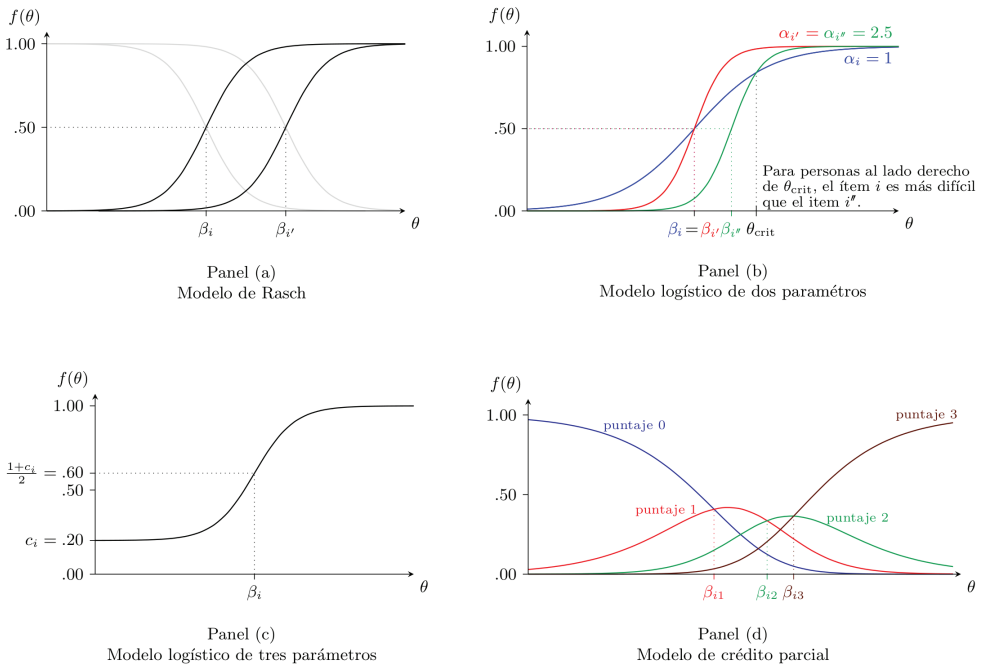
En esta ecuación $\exp(x)$ significa e^x (con e la base de los logaritmos naturales, e ≈ 2.71828). Como comentamos anteriormente, θ_p indica el nivel o el dominio de la persona p en el rasgo latente; β_i, por otro lado, se conoce como la dificultad del ítem i y corresponde con el nivel requerido en el rasgo latente para que un individuo tenga una probabilidad del 50% de dar la respuesta correcta al ítem i. (Nótese que, cuando $\theta_p = \beta_i$ en la Ecuación 2a, entonces la probabilidad de acertar el ítem es igual a 0.50.) Puesto que los ítems en el modelo de Rasch son dicotómicos, la probabilidad de que la persona falle el ítem se da por:

$$\begin{aligned} \Pr(Y_{pi} = 0) &= 1 - \Pr(Y_{pi} = 1) \\ &= 1 - \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \\ &= \frac{1}{1 + \exp(\theta_p - \beta_i)}. \end{aligned} \tag{2b}$$

El panel (a) de la Figura 1 muestra (en color negro) la función característica de dos ítems en el modelo de Rasch. Los ítems se colocan en la dimensión latente con base en su parámetro de dificultad, justamente –como comentamos– donde la función característica indica una

probabilidad de 50%. Nótese que el ítem *i* es más fácil que el ítem *i'* ($\beta_i < \beta_{i'}$) y que, efectivamente, para cualquier nivel de la persona en el rasgo latente, la probabilidad de acertar es mayor para el ítem *i* en comparación con el ítem *i'*. Recuérdese que la función característica de ítems dicotómicos indica la probabilidad de responder en la categoría "respuesta correcta" (dada por la Ecuación 2a); aunque normalmente no se muestra la función característica para la categoría "incorrecta", en el panel (a) de la Figura 1, incluimos para fines didácticos estas funciones características (correspondientes a la Ecuación 2b) en color gris ligero.

Figura 1. Curvas características en distintos modelos de respuesta al ítem



El Modelo Logístico de Dos Parámetros

Birnbaum (1968) añadió al modelo de Rasch un parámetro más para cada ítem. Este parámetro, comúnmente denotado α_i (un número real positivo), se suele interpretar como la discriminación del ítem *i*. La ecuación básica del modelo logístico de dos parámetros (2PLM, por sus siglas en inglés) es la siguiente:

$$\Pr(Y_{pi} = 1) = \frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]}. \qquad (3)$$

Nótese que, si el parámetro de discriminación α_i es igual a uno para todos los ítems, el modelo 2PL se simplifica y se obtiene el modelo de Rasch. El panel (b) de la Figura 1 mues-

tra la función característica de tres ítems en el 2PLM. Se observa que los ítems con un grado de discriminación más alto (por ejemplo, el ítem de color más oscuro) tienen la pendiente más pronunciada. Efectivamente, es lo que significa la discriminación de un ítem en este modelo: para dos personas (con posiciones cercanas al grado de dificultad del ítem) se hace más grande la diferencia entre sus probabilidades de acertar conforme el ítem discrimina más. En el caso más extremo (cuando $\alpha_i \to \infty$), la discriminación es máxima: personas con un nivel de habilidad menor que la dificultad β_i del ítem tienen probabilidad 0 de acertar, mientras que los que tienen el nivel arriba de β_i tienen probabilidad 1.

La gráfica muestra adicionalmente una diferencia importante entre el 2PLM y el modelo de Rasch: las curvas características en el 2PLM generalmente se cruzan, lo cual nunca puede ocurrir en el modelo de Rasch. Como consecuencia de esta característica, la interpretación del parámetro de dificultad ya no es uniforme para todas las personas. En la gráfica se observa, por ejemplo, que $\beta_i < \beta_{i''}$, mientras que para las personas con un nivel mayor que θ_{crit} (indicado en la gráfica), el ítem i'' es más fácil que el ítem i. En general, en modelos TRI más complejos (con más parámetros), la interpretación de los parámetros suele ser menos unívoca.

La gráfica muestra también una similitud entre el modelo de Rasch y el 2PLM: para personas con un nivel θ_p muy bajo en el rasgo latente, la probabilidad de acertar el ítem se acerca a 0; en contraste, las personas con la θ_p muy alta tienen probabilidades cercanas a 1. El siguiente modelo permite curvas características donde la probabilidad no se acerca a 0 para niveles muy bajos en el constructo latente.

El Modelo Logístico de Tres Parámetros

Los modelos anteriores pueden presentar una inconveniencia cuando se utilizan para analizar exámenes de opción múltiple. En un examen de este tipo es poco plausible que la probabilidad de acertar se acerque a 0, incluso para personas con un dominio muy bajo en el rasgo latente, ya que se puede responder correctamente por adivinación (azar). Es por esta razón que Birnbaum (1968) propuso el modelo logístico de tres parámetros (3PLM). En este modelo, además de los parámetros α_i y β_i, cada ítem tiene un tercer parámetro, c_i, el cual se denomina "parámetro de adivinación". Este parámetro (cuyo valor teóricamente puede estar entre 0 y 1) se suele interpretar como la probabilidad de acertar el ítem en caso de que una persona no sepa la respuesta. La ecuación básica que permite calcular la probabilidad de acertar en el 3PLM se da por:

$$\Pr(Y_{pi} = 1) = c_i + (1 - c_i)\,\frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]}.$$

Es claro que este modelo generaliza el 2PLM: cuando en la ecuación anterior $c_i = 0$ para todos los ítems, se obtiene la Ecuación 3. El panel (c) de la Figura 1 muestra como ejemplo

la curva característica de un reactivo en el 3PLM. Se observa que, efectivamente, la curva no baja hasta una probabilidad de 0 para las personas con el nivel θ_p muy bajo, sino a la probabilidad de $c_i = 0.20$ (en este ejemplo). Además, se observa que el parámetro β_i, contrario a los modelos anteriores, ya no corresponde con el nivel en el rasgo latente donde la probabilidad de acertar el ítem es igual a 50% (sino donde la probabilidad es igual a $(1 + c_i) / 2$).

Para una interpretación correcta del parámetro c_i en el 3PLM son útiles las siguientes dos reflexiones. Primero, cada ítem tiene un parámetro c_i y distintos ítems suelen tener distintos valores para c_i. A veces surge una confusión porque, de ser interpretado como la probabilidad de elegir por azar la opción correcta en un reactivo de opción múltiple, el parámetro c_i debería tener necesariamente el valor $1/m$, donde m es el número de opciones de respuesta para el reactivo (por ejemplo, $c_i = 0.25$ si el reactivo tiene cuatro opciones de respuesta). Si bien este razonamiento sería correcto si una persona respondiera el reactivo *ciegamente* (como si no hubiera leído las opciones), el 3PLM toma en cuenta que la opción correcta puede ser más o menos atractiva para personas con un nivel muy bajo en el rasgo latente, por ejemplo, porque estas personas se dejan atraer más por algún distractor o, al contrario, algún distractor es obviamente incorrecto, incluso para personas "ignorantes". Segundo, es importante tener claro que c_i es un parámetro asociado con los ítems y, en este sentido, hace referencia a una propiedad de los ítems; esto implica que, si cierto reactivo de opción múltiple tiene un valor alto para c_i, entonces la opción correcta de este reactivo es relativamente atractiva para *todas* las personas con nivel bajo en el rasgo latente. Específicamente con respecto a los exámenes de opción múltiple se ha introducido el rasgo de *testwiseness*, lo cual se define como la habilidad que pueden tener ciertas personas para aprovechar aspectos del formato y de la redacción de los reactivos para aumentar su calificación en el examen. Sin embargo, esta habilidad, al ser una característica de las personas (ajena al constructo que con el examen se quiere medir), no puede ser captada por el parámetro c_i en el 3PLM. (Se podría tomar en cuenta el rasgo de *testwiseness* en un modelo TRI; este sería necesariamente un modelo multidimensional, ya que, además del rasgo latente que se quiere medir, incluiría también el *testwiseness*.)

El Modelo de Crédito Parcial

Si las respuestas en las preguntas de un examen no son binarias (correcto-incorrecto) o no se desea dicotomizarlas para que así lo sean, los tres modelos previos no sirven para su análisis. En este caso hay que recurrir a modelos TRI para ítems politómicos. Aquí describimos brevemente uno de estos modelos: el modelo de crédito parcial (PCM, por sus siglas en inglés) de Masters (1982, 2016).

El PCM considera, de manera general, el caso de un examen donde cada pregunta se califica con un número entero de 0 a m_i (con $m_i \geq 1$; se permite que la puntuación máxima es diferente para distintos ítems, por eso se añadió el índice i en m_i), donde una puntuación mayor es indicador de un nivel más alto en el (único) rasgo latente que el examen quiere medir. Por un lado, este modelo, igual que los anteriores, asigna un parámetro, θ_p, a cada persona; por otro lado, a cada ítem se asignan m_i parámetros (un parámetro para cada

categoría/puntaje mayor que 0): β_{i1}, β_{i2}, ..., β_{im}. La ecuación básica, que no presentamos aquí por su complejidad matemática (pero se puede consultar en las publicaciones mencionadas de Masters), permite calcular, con base en estos parámetros, la probabilidad de que la persona p en el ítem i obtenga un puntaje igual a j (para cualquier valor de $j = 0,1, ..., m_i$).

El panel (d) de la Figura 1 muestra un ejemplo de un ítem en el PCM. Debido a que se trata de un ítem politómico con puntaje máximo de $m_i = 3$, la gráfica muestra cuatro curvas características distintas (asociadas con los puntajes 0, 1, 2 y 3). Nótese que las curvas características asociadas con el puntaje mínimo de 0 y el puntaje máximo de 3 son monótonamente decrecientes y crecientes, respectivamente, mientras que las categorías intermedias tienen curvas unimodales.[2] Esta forma de las curvas no sorprende: efectivamente, conforme el nivel de la persona en el rasgo latente aumenta, la probabilidad del puntaje de 0 baja y la del puntaje máximo crece. Por otro lado, con respecto a los puntajes intermedios, ni las personas con un nivel extremadamente bajo ni con un nivel extremadamente alto tendrán estos puntajes; las primeras porque no logran superar el puntaje de 0, las últimas porque logran tener el puntaje máximo de 3.

La gráfica nos muestra también cómo se pueden interpretar los parámetros β_{i1}, β_{i2} y β_{i3} de este ítem (y, en general, de los ítems en el PCM). Se observa que β_{ij} corresponde con la posición en el rasgo latente donde las curvas características asociadas con los puntajes j y $j - 1$ se cruzan. Por ejemplo, β_{i1} indica el nivel en el rasgo latente desde el cual es más probable tener un puntaje de 1 que un puntaje de 0, y en este sentido se puede interpretar como la dificultad de la categoría 1 relativa a la categoría 0. Nótese que la misma interpretación aplica al único parámetro (β_i) del ítem en el modelo de Rasch, que indica el nivel en el rasgo latente donde acertar y fallar tienen la misma probabilidad. En efecto, el PCM es una generalización para ítems politómicos del modelo de Rasch y cuando $m_i = 1$, el PCM es idéntico al modelo de Rasch. Cabe mencionar que en desarrollos posteriores del PCM, se incluyó también un parámetro de discriminación (α_i) para cada ítem (similar al 2PLM), lo cual llevó al modelo de crédito parcial generalizado.

CONSIDERACIONES RELEVANTES EN LA APLICACIÓN DE UN MODELO TRI

Hasta este punto, hemos dado una introducción teórica a los modelos TRI. A fin y al cabo, los modelos estadísticos son una teoría, formalizada en lenguaje matemático, para describir un fenómeno; en el caso de los modelos psicométricos de la TRI, son una teoría sobre la conducta de personas que se enfrentan a (los reactivos en) un examen. Sin embargo, para que los

[2] Al inicio de esta sección, asociamos las curvas características unimodales con los modelos de proximidad o punto ideal. Para una comprensión correcta, estos modelos especifican curvas características unimodales, no solo para las posibles categorías intermedias (en caso de ítems politómicos), sino también para la categoría más alta. El PCM no es un modelo de proximidad (sino de dominancia) ya que la curva característica asociada con el puntaje máximo es creciente (o, de manera equivalente: conforme el nivel en el rasgo latente aumenta, el puntaje esperado en el ítem aumenta también).

modelos tengan relevancia práctica es importante concretar su aplicación a datos observados. En general, tal aplicación implica resolver las siguientes tres cuestiones: (1) ¿Qué estrategia se utilizará para asignar valores a los parámetros incógnitos del modelo? Este es el problema que resuelve la estimación de parámetros. (2) Una vez que se han estimado los valores de los parámetros, ¿se puede concluir que el modelo da una descripción adecuada de los datos observados? Esta cuestión se refiere a la bondad de ajuste del modelo. (3) Tomando en cuenta que en las aplicaciones prácticas siempre hay factores no deseados que distorsionan los datos observados, ¿qué tan confiables son los resultados (es decir, las estimaciones de los parámetros) obtenidos mediante el modelo? Esta cuestión hace referencia a la precisión de las estimaciones. A continuación, exponemos brevemente cómo estas cuestiones se suelen abordar en el contexto de los modelos TRI.

Estimación de Parámetros

En el caso de la TRI, los datos observados a los que se aplican los modelos son las respuestas de una muestra de personas al conjunto de reactivos que conforman el examen. En la sección anterior presentamos los principios generales de los modelos TRI y explicamos de manera breve algunos de estos modelos que son muy utilizados en la práctica. Aclaramos que los modelos especifican parámetros (valores numéricos con un significado particular en el contexto de un modelo concreto), para personas e ítems, pero no mencionamos cómo se obtienen estos valores numéricos. Un primer paso en la aplicación de un modelo TRI es el ajuste del modelo a datos observados, es decir, la estimación de sus parámetros con base en estos datos.

Consideremos como ejemplo una aplicación del modelo de Rasch a los datos que se presentan (parcialmente) en la Tabla 1; son las respuestas (1 = correcto; 0 = incorrecto) de 500 personas a 60 reactivos. Aplicar el modelo de Rasch a estos datos implica estimar los 500 parámetros de personas y los 60 parámetros de ítems. En la Tabla 1 hemos añadido el acento ^ al parámetro (por ejemplo: $\hat{\theta}_p$ y $\hat{\beta}_i$) en la última columna y fila, para diferenciar el valor estimado con base en datos observados del valor teórico-verdadero en el modelo.

Existen diferentes métodos para estimar los parámetros de un modelo estadístico. Para los modelos TRI, el método más utilizado es la estimación por máxima verosimilitud (MLE, por sus siglas en inglés) y es la estrategia comúnmente implementada en los algoritmos de estimación en *software* comercial como BILOG, MULTILOG, PARSCALE (*Scientific Software International Inc.*) y WinSteps (*Winsteps Inc.*) y *software* libre como MIRT (*Educational Testing Service*), jMetrik (Universidad de Virginia) y los paquetes especializados de R. En el marco de este capítulo introductorio no tenemos el espacio para indagar sobre la MLE. Solo mencionamos que existen distintas variantes de MLE y que cada variante tiene sus ventajas y desventajas: el método por máxima verosimilitud marginal es muy flexible, pero añade un supuesto al modelo (el supuesto de que los parámetros θ_p se extraen de manera aleatoria de una distribución normal); el método por máxima verosimilitud condicional no requiere supuestos adicionales, pero solo es aplicable a un subgrupo de modelos (que incluyen la propiedad de que el puntaje total de una persona en el test es suficiente para estimar el

parámetro θ_p de esta persona); y el método de máxima verosimilitud conjunta, aunque generalmente aplicable, tiene la desventaja teórica de que las estimaciones no son consistentes. Cabe mencionar que, aunque (las variantes de) MLE sigue siendo el procedimiento estándar de estimación para modelos TRI, cada vez es más común realizar estimaciones en el marco bayesiano.

Evaluación de la Bondad de Ajuste del Modelo

Los métodos de estimación asignan valores numéricos a los parámetros de personas e ítems que, en cierto sentido, son los mejores. Sin embargo, incluso con las mejores estimaciones puede ser que el modelo no describa de manera aceptable los datos observados; en concreto, puede ser que las probabilidades derivadas a partir de la ecuación básica del modelo no tengan una correspondencia adecuada con las respuestas de las personas a los reactivos. Por ejemplo, considérese en una aplicación del modelo de Rasch a una persona cuyo nivel en el rasgo latente se estimó como intermedio, mientras que los datos muestran que esta persona acierta los ítems difíciles (con altos valores para $\hat{\beta}_i$) y falla los ítems fáciles (con valores bajos para $\hat{\beta}_i$). Esta situación señala una discrepancia entre los datos y el modelo, ya que las probabilidades de acertar según el modelo son altas justamente donde la persona falla y, al contrario, son bajas cuando la persona acierta. Aunque el modelo no excluye la posibilidad de que ocurran estas discrepancias, si en los datos ocurren frecuentemente, lo más apropiado entonces sería rechazar el modelo y buscar otro modelo TRI que dé una mejor explicación a los datos observados.

Un mal ajuste del modelo a los datos apunta a que uno o más supuestos se están incumpliendo. Existen muchas estrategias para evaluar la bondad de ajuste de un modelo TRI. Estas estrategias pueden evaluar el ajuste de manera global o bien, pueden detectar incumplimientos particulares, quizás asociados con ciertas personas o ítems, o con supuestos específicos (como unidimensionalidad o independencia local). A continuación, describimos brevemente los principios de dos grupos de métodos de evaluación de la bondad de ajuste para modelos TRI.

- *Métodos gráficos/visuales*
  Este método (más útil para modelos unidimensionales) evalúa la bondad de ajuste de los distintos ítems a través de una gráfica que permite comparar (las probabilidades de responder en cierta categoría mostradas por) la curva característica de cada ítem con la proporción de respuestas en distintos subgrupos de personas que tienen un nivel estimado similar en el rasgo latente. Para ilustrarlo, volvamos a considerar los datos en la Tabla 1 y la aplicación del modelo de Rasch a estos datos. Como siguiente paso, dividimos el grupo total de 500 personas en 10 grupos de aproximadamente 50 personas que tienen estimaciones $\hat{\theta}_p$ similares (es decir, un primer grupo con las 50 personas con las $\hat{\theta}_p$ más bajas, después un segundo grupo con las siguientes 50 personas, etcétera.) y calculamos en cada grupo la proporción de aciertos en el ítem *i* bajo consideración. En la gráfica de la curva característica se añade un punto a la altura de la proporción calculada

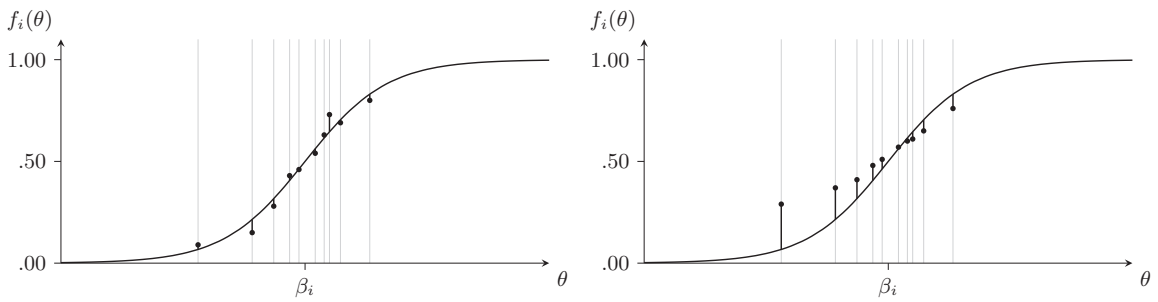
para cada grupo (donde el grupo se posiciona en la abscisa según la media de las $\hat{\theta}_p$). El panel (a) de la Figura 2 muestra un ejemplo de un ítem con buen ajuste: las proporciones de aciertos en los distintos grupos se acercan a (las probabilidades teóricas en) la curva; el panel (b) muestro un ítem con mal ajuste, donde las discrepancias entre las proporciones y las probabilidades son más grandes.

Tabla 1. Ejemplo de datos observados y los parámetros asociados en el modelo de Rasch.

		Ítems									
		1	2	3	4	⋯	*i*	⋯	60		
Personas	1	1	0	0	1	⋯	1	⋯	1	→	$\hat{\theta}_1$
	2	0	1	0	1	⋯	1	⋯	0	→	$\hat{\theta}_2$
	3	0	0	1	1	⋯	1	⋯	0	→	$\hat{\theta}_3$
	4	0	1	0	0	⋯	0	⋯	0	→	$\hat{\theta}_4$
	5	0	0	0	1	⋯	0	⋯	1	→	$\hat{\theta}_5$
	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋱	⋮		⋮
	p	1	1	0	0	⋯	1	⋯	1	→	$\hat{\theta}_p$
	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋱	⋮		⋮
	500	1	0	0	1	⋯	0	⋯	0	→	$\hat{\theta}_{500}$
		↓	↓	↓	↓		↓		↓		
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	⋯	$\hat{\beta}_i$	⋯	$\hat{\beta}_{60}$		

Nota. La última columna y la última fila indican los parámetros estimados de personas e ítems, respectivamente, al aplicar el modelo de Rasch a estos datos.

Figura 2. Evaluación de la bondad de ajuste de dos ítems en el modelo Rasch (método gráfico)



Panel (a)
Ítem con buen ajuste

Panel (b)
Ítem con mal ajuste

- *Índices de bondad de ajuste*
  Los programas informáticos para aplicar modelos TRI permiten calcular una amplia gama de índices de bondad de ajuste. Estos índices cuantifican de cierta manera la diferencia entre los datos observados y los datos esperados según el modelo (similar al ejemplo que se acaba de describir en el método gráfico, donde de manera visual se aprecia la diferencia entre las proporciones de aciertos observadas en la muestra y las probabilidades teóricas, que se pueden interpretar como proporciones esperadas según el modelo). Comúnmente se distingue entre índices exactos, que permiten realizar una prueba exacta de bondad de ajuste del modelo a los datos (donde un ejemplo clásico es el estadístico ji-cuadrado) e índices de ajuste aproximados. Estos últimos se construyeron a partir de la idea que cualquier modelo es equivocado en el sentido de que no puede abstraer la realidad con todos sus detalles y, por lo tanto, se rechazará cuando hay suficientes datos disponibles. Sin embargo, al proponer un modelo se espera poder usarlo para obtener información y conocimiento conciso acerca de los datos observados y es suficiente con que el modelo tenga un ajuste suficientemente bueno, aunque no sea perfecto. Un índice de bondad de ajuste aproximado muy utilizado es el RMSEA, que con valores más bajos indica un mejor ajuste. Para este tipo de índices, la literatura presenta puntos de corte a partir de los cuales se considera que el ajuste es "aceptable" o "bueno" (por ejemplo, RMSEA < .08 se considera un ajuste aceptable; RMSEA < .05 es excelente).

Antes de concluir esta sección sobre bondad de ajuste, queremos llamar la atención a un incumplimiento de los supuestos en modelos TRI que en los últimos años ha recibido más atención y cuya detección en exámenes de alto impacto se considera esencial. Se trata de *funcionamiento diferencial del ítem* (DIF). Uno de los supuestos clave en los modelos TRI es que los parámetros de los ítems tengan el mismo valor (técnicamente se dice que son "invariantes") para cualquier muestra de personas extraída de la población para la cual el modelo se cumple (y, de forma similar, que los parámetros de las personas son invariantes para dis-

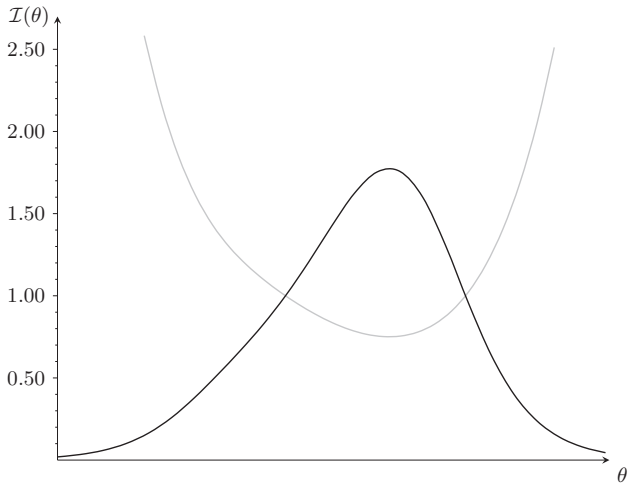
tintas submuestras de ítems). Se presenta DIF si los ítems se comportan de manera diferente en distintas subpoblaciones, por ejemplo, con estudiantes de familias con recursos limitados o en la subpoblación de mujeres, etc. Cuando se presenta funcionamiento diferencial, las comparaciones de distintas subpoblaciones pueden resultar inapropiadas (en el sentido de que una $\hat{\theta}_p$ más baja en un grupo no necesariamente corresponde con un nivel más bajo en el rasgo latente que se quiere medir). Existe mucha literatura que presenta métodos para evaluar y detectar DIF en pruebas analizadas con modelos TRI.

Confiabilidad (Precisión de las Estimaciones)

En la sección sobre estimación de parámetros, enfatizamos la diferencia conceptual entre los valores teóricos para los parámetros (θ_p, β_i, α_i, etc.) y las estimaciones correspondientes ($\hat{\theta}_p$, $\hat{\beta}_i$, $\hat{\alpha}_i$, etc.). Es de esperar que, por la influencia de factores aleatorios, los valores estimados no coincidan perfectamente con los teóricos (los cuales, en principio, son siempre desconocidos). De ahí surge el tema de qué tanto difieren las estimaciones de los valores verdaderos. Aunque nunca sabremos para un caso concreto cuál es la diferencia entre, digamos, θ_p y $\hat{\theta}_p$, sí es posible hacer afirmaciones generales sobre la precisión de las estimaciones (por ejemplo, cuál es la diferencia esperada –o promedio– entre la estimación y el valor teórico-verdadero correspondiente).

En la teoría clásica de los tests, la precisión de la medición depende de la confiabilidad del test. En particular, la confiabilidad permite calcular el error estándar de medición (el cual es la desviación estándar del error de medición, y en este sentido, se interpreta como la diferencia esperada entre la puntuación verdadera y la puntuación observada en un test). En la TRI, se toma otro enfoque con respecto al error de medición, el cual dependerá de la estrategia de estimación particular que se utilice. En el caso de MLE, por ejemplo, la precisión de la medición (es decir, la diferencia esperada entre el valor estimado y el valor teórico-verdadero del parámetro) se cuantifica a través de la función de información de Fisher. Cuando se aplica a la estimación del parámetro θ_p, la información de Fisher tiene un papel similar a la confiabilidad en la TCT, en el sentido de que permite calcular el error estándar de medición (en particular, el error estándar de medición es el inverso de la raíz cuadrada de la información). Una diferencia fundamental entre la información de Fisher y la confiabilidad (o entre el error estándar de medición en la TRI *versus* la TCT) es que la información depende (del valor del parámetro) de la persona (véase la Figura 3), mientras que la confiabilidad es constante en una población de personas. Es decir, en la TRI las $\hat{\theta}_p$ tienen una precisión distinta para diferentes personas que contestan el test. Intuitivamente, no resulta extraño que la precisión de las estimaciones del nivel en el rasgo latente sea baja para personas con calificaciones muy altas o muy bajas. (Si necesitáramos una precisión más alta para discriminar entre sustentantes en los extremos del rasgo latente, sería oportuno realizar otras pruebas, por ejemplo, un test adaptativo que ajuste automáticamente el nivel de dificultad de los ítems que se presentan al sustentante conforme avanza en la prueba.) Cabe mencionar que también en un marco bayesiano existen métodos para cuantificar la precisión de las estimaciones, la cual puede variar entre distintas personas.

Figura 3. Ejemplo de la función de información (negro) y el error estándar de medición (gris)



DISCUSIÓN Y CONCLUSIONES

En este capítulo hemos dado una introducción a la TRI dirigida a profesores sin antecedentes técnicos en psicometría, con una descripción breve de algunos de los modelos más utilizados y temas relevantes para su aplicación en el análisis de exámenes. Debido a sus fundamentos arraigados en la estadística matemática, publicaciones sobre la TRI suelen estar intercaladas con fórmulas complejas y lenguaje técnico difíciles de entender por usuarios no expertos en este tema. Esto indudablemente es la razón principal por la que el uso de modelos TRI para el análisis de exámenes, a pesar de sus numerosas ventajas sobre la TCT, sigue siendo relativamente marginal.

Otro inconveniente de la TRI, que muchos autores mencionan para explicar su uso poco frecuente en la práctica, es la necesidad de muestras relativamente grandes para la aplicación de los modelos. Incluso para modelos relativamente sencillos (como los presentados en este capítulo) se requieren como mínimo unos cuantos cientos de personas para obtener estimaciones suficientemente precisas. En este sentido, es complicado aplicar los modelos TRI para exámenes que se aplican a grupos de 60 personas, como es el caso para muchos profesores en instituciones educativas.

Concluimos este capítulo mencionando dos de las ventajas más importantes de la TRI sobre la TCT. Aunque más de índole teórico, los autores consideramos que la ventaja principal de la TRI es que da un fundamento robusto para la medición del (los) constructo(s) latente(s) que se quiere(n) medir con el examen. Mientras que los modelos TRI describen explícitamente (de manera probabilística, a través de la ecuación básica del modelo) cómo las respuestas en el examen dependen del nivel de la persona en el (los) constructo(s) latente(s), la teoría clásica ni siquiera hace referencia a un constructo latente en el desarrollo del modelo. (Nótese, al respecto, que la puntuación verdadera en la TCT no solo recoge el efecto

del (los) rasgo(s) latente(s) que se quiere medir con el test, sino también de *todos* los factores que ejercen una influencia sistemática entre las distintas réplicas conceptuales en las que se basa la definición de la puntuación verdadera y del error.) La TRI permite llegar a conclusiones sobre la dimensionalidad del rasgo latente (¿cuántos rasgos latentes subyacen al test y cómo se relacionan entre sí?) y, además, cómo esta(s) dimensión(es) se relaciona(n) con los reactivos en el examen.

La segunda ventaja importante se deriva de la propiedad, anteriormente mencionada, de la invarianza de los parámetros del modelo. Aunque son importantes las consideraciones sobre el diseño para la recopilación de los datos y el método de estimación (véase Hambleton, 1994), esta invarianza generalmente ofrece una base sólida para comparar los parámetros de diferentes personas, aunque respondieran distintos conjuntos de ítems y, viceversa, para comparar los parámetros de diferentes ítems, aunque hayan sido respondidos por distintos grupos de personas. Esta propiedad conlleva ventajas importantes cuando, es imposible (o al menos no deseable) aplicar el mismo examen a todas las personas, por ejemplo, cuando un examen se aplica en diferentes días a diferentes grupos y existe el riesgo que los reactivos sean compartidos entre los sustentantes. En estos casos, un análisis con base en la TCT no permitiría diferenciar entre puntuaciones más altas porque la variante del examen aplicada en cierto día era más fácil o bien porque los sustentantes de ese día tenían un nivel más alto en el (los) rasgo(s) evaluado(s). Cabe mencionar que la explotación de la propiedad de invarianza se lleva a un extremo en el caso de los llamados "tests adaptativos informatizados", que presentan, a través de una computadora, a cada persona un conjunto de reactivos personalizado mientras que los niveles estimados de distintas personas en el (los) rasgo(s) latente(s) son directamente comparables. El uso de tests adaptativos informatizados lleva a exámenes más cortos y estimaciones más precisas, aún para sustentantes con valores extremos en el (los) rasgo(s) latente(s), debido a que cada persona responde a ítems ajustados a su nivel. Este tipo de pruebas, que hacen uso exquisito de la tecnología, no son compatibles con la TCT, pero sí con la TRI.

Agradecimiento

Los autores agradecen a Georgina García Rodríguez por la lectura crítica y comentarios realizados sobre una versión previa de este capítulo.

REFERENCIAS

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord y M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 396–479). Addison-Wesley.

Hambleton, R. K. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema*, *6*(3), 535–556.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Psychometric Corporation. https://www.psychometricsociety.org/sites/main/files/file-attachments/mn07.pdf

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.
Masters, G. N. (2016). Partial credit model. En W. J. van der Linden (Ed.). *Handbook of item response theory* (Vol. 1, pp. 109–126). CRC Press, Taylor & Francis.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.
Reckase, M. D. (2009). *Multidimensional item response theory.* Springer.
Thissen, D., y Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.
van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory* (Vol. 1). CRC Press, Taylor & Francis.