4

Introducción a la Teoría de la Respuesta al Ítem

Introducción

La Teoría Clásica de los Tests (TCT) sigue siendo el modelo predominante para la construcción de tests psicológicos tanto por la sencillez de sus procedimientos y supuestos como por su demostrada utilidad práctica. Sin embargo, se conocen bien las limitaciones teóricas del modelo clásico y se ha desarrollado un nuevo enfoque psicométrico, la Teoría de la Respuesta al Ítem (TRI), que permite superarlas. La TRI supone una aproximación más fina en el estudio de las propiedades psicométricas de un test, ya que modela de forma más realista las respuestas de las personas, toma los ítems como unidad de análisis y permite describir algunas propiedades psicométricas del instrumento mediante indicadores que son *invariantes* y que no dependen de la muestra en la que se aplique (siempre que se cumplan una serie de supuestos).

Los principios de la TRI se remontan a los trabajos de Thurstone (1925; 1927), Lawley (1943), Guttman (1944) y Lazarsfeld (1950; 1959). El interés era obtener instrumentos de medida cuyas propiedades no dependieran de la muestra en la que se aplicara. Fue Lord (1952) quien hizo la aportación definitiva, presentando el primer modelo de TRI en un monográfico de la revista *Psychometrika*. Sin embargo, es después de los 60 cuando se empiezan a desarrollar nuevos modelos y procedimientos que facilitan su aplicabilidad (Rasch, 1960; Lord y Novick, 1968). En los 80, será Lord el que acuña la denominación de Teoría de la Respuesta al Ítem (Lord, 1980) y se publican diversos libros que constituyen las obras de referencia sobre la TRI (Hambleton y Swaminathan, 1985; Hambleton, Swaminathan y Rogers, 1991; Hulin, Drasgow y Parsons, 1983; Lord, 1980). Después de los 80, y sobre todo en las dos últimas décadas, no han dejado de proponerse nuevos modelos y aplicaciones de la TRI. Descripciones de los viejos y nuevos modelos pueden encontrarse en numerosas fuentes, en inglés (De Ayala, 2009; Embretson y

Hershberger, 1999; Embretson y Reise, 2000; Ostini y Nering, 2006; Thissen y Wainer, 2001; Van der Linden y Hambleton, 1997) y en castellano (Lopez-Pina, 1995; Martinez Arias et al., 2006; Muñiz, 1996, 1997; Revuelta, Abad y Ponsoda, 2006).

Hasta hace poco, en España, existían pocos instrumentos psicológicos desarrollados exclusivamente en el marco de la TRI. Una razón es que los nuevos procedimientos de la TRI son más costosos, complejos y exigentes en las comprobaciones de los supuestos. Sin embargo, en otros países, su uso es muy frecuente en pruebas de selección o acreditación aplicadas a grandes muestras (p.ej., SAT, GRE, TOEFL, ASVAB, etc.) y en contextos de evaluación educativa. La TRI se ha convertido también en una herramienta indispensable cuando se quiere comparar las diferencias entre diversos países o culturas (p.ej., en los tests que forman el TIMSS o las pruebas educativas del proyecto OECD/PISA). En España, son cada vez más los tests basados específicamente en la TRI o en los que se complementa el estudio psicométrico realizado desde la TCT. También son cada vez más frecuentes los trabajos de investigación aplicada sustentados en estos modelos y el desarrollo de ciertos tipos de tests que requieren de la TRI, como son los Tests Adaptativos Informatizados, de los cuales existen ya diversas versiones operativas para evaluar diferentes atributos.

El presente capítulo representa una introducción al tema, incluyendo únicamente aspectos generales y los modelos de TRI para ítems de rendimiento óptimo, donde las respuestas son cuantificadas como acierto o error. En los capítulos 11 y 12 se profundizará en los procedimientos estadísticos de estimación y ajuste, así como en otro tipo de modelos. En los capítulos 13 y 15 se describen algunas de las principales aplicaciones.

Limitaciones de la TCT

Existen diversas razones por las que la TRI supone un modelo teóricamente más adecuado que la TCT. Entre las limitaciones de la TCT, superables desde la TRI, cabe destacar:

1. *Ausencia de invarianza de los parámetros.* En la TCT la puntuación verdadera V_i es un parámetro de la persona cuyo valor depende del conjunto particular de ítems administrados. Es claro que distintos tests, con distinta longitud o distinta dificultad, darán lugar a distinta puntuación verdadera para la misma persona. No parece razonable un modelo en el que la puntuación verdadera de la persona depende de la versión del test que apliquemos (p.ej., fácil o difícil). Por otro lado, las propiedades psicométricas de los ítems (su media, su varianza, su índice de discriminación, etc.) también dependen de la distribución del rasgo en la muestra donde se obtienen. La TCT no proporciona un modo sencillo de saber cuál sería la dificultad de un ítem en otra muestra distinta a la que se ha aplicado el test. Frente a la TCT, una de las propiedades de la TRI es que los parámetros estimados son invariantes si se cumplen los supuestos del modelo; de esta manera, en la TRI, el valor del parámetro que indica el verdadero nivel de rasgo de un evaluado no depende de los ítems aplicados (p.ej., si son fáciles o difíciles). Asimismo, el valor de los parámetros de los ítems no depende de la muestra donde se obtienen (p.ej., si es de alto o bajo nivel de habilidad).
2. *Se asume que la precisión del test es la misma, independientemente del nivel de rasgo medido.* Sin embargo, un test puede ser más o menos preciso para un nivel de rasgo en

función, por ejemplo, de la dificultad de los ítems aplicados; si los ítems son difíciles, el test discriminará mejor en los niveles altos. La TRI permite obtener la precisión con la que cada persona es medida, según su nivel de rasgo y en función de los ítems concretos que se le hayan aplicado.

3. *No se dispone de indicadores de bondad de ajuste que nos informen del grado en que el modelo se ajusta a los datos.* Los supuestos de paralelismo son los únicos contrastables empíricamente (ver capítulo 10), pero raramente se realizan estas comprobaciones pues, por un lado, requieren la elaboración de distintas formas del test y, por otro, se sabe que el supuesto de paralelismo estricto raramente se cumple, dado que es muy difícil elaborar tests que sean realmente paralelos. En la TRI se dispone de indicadores de bondad de ajuste que permiten estudiar el grado en que los datos se ajustan al modelo y a los supuestos establecidos.

La TRI permite superar varias de las limitaciones de la TCT mediante una metodología más compleja, que requiere establecer modelos matemáticos, la estimación de sus parámetros y enjuiciar el ajuste entre datos y modelos.

La Curva Característica del Ítem

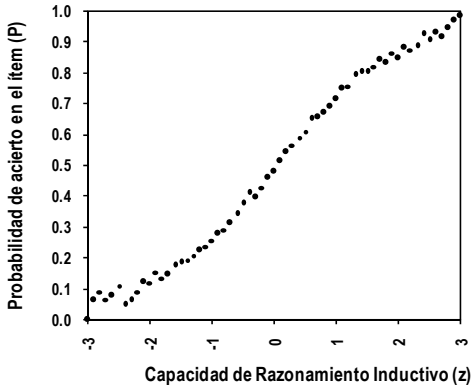
Para comenzar a resolver las limitaciones de la TCT, lo primero que se necesita es un modelo que nos indique cómo varía la dificultad de un ítem en función del nivel de rasgo. Para ello, desde la TRI se ha desarrollado el concepto de *Curva Característica del Ítem* (*CCI*). Una CCI nos indica la probabilidad que tienen de acertar un ítem las personas que lo responden. Esta probabilidad depende, naturalmente, del nivel de la persona en la variable medida, pero también de las características del ítem en concreto.

Podemos ver esto mediante un ejemplo. Supongamos que tenemos un test largo que mide Capacidad de Razonamiento Inductivo y que ha sido aplicado a una muestra numerosa de personas (100.000, por ejemplo). Supongamos que la menor y mayor puntuación obtenidas en el test son 50 y 150 y que la puntuación en el test sea un buen indicador del nivel de rasgo verdadero. Para trabajar en una escala de interpretación más clara, utilizaremos la puntuación en el test en puntuaciones típicas (–3 indica una puntuación baja, 0 una puntuación media y 3 una puntuación alta). Vamos a representar el rendimiento en un ítem concreto de la siguiente forma: Nos fijamos en todas las personas que han obtenido la puntuación en torno a –3 (supongamos que son 132) y vemos cuantas han acertado el ítem (supongamos que han sido sólo 5) y calculamos la proporción (5/132 = 0,04). Hacemos lo mismo con los que obtuvieron en el test una puntuación en torno a –2,9 puntos (y obtenemos la proporción, supongamos que 0,15),... con las que obtuvieron en el test puntuación en torno a 0.0 (la proporción fue 0,48),... con las que obtuvieron puntuación en torno a 3 (la proporción fue 0,98). La Figura 4.1 muestra la proporción de aciertos en el grupo de personas que obtuvo en el test puntuaciones en torno a –3, –2,9, –2,8, etc.

Puede verse que cuanto mayor es la puntuación en el test, mayor es la proporción de aciertos en el ítem. A una puntuación de 0 le corresponde una proporción de 0,48, lo que indica que para personas con ese nivel de rendimiento en el test resultará un ítem de dificultad intermedia; mientras que a una puntuación de 3,0 le corresponde una proporción de

0,98 (el ítem resultará fácil para ese nivel). La función de la Figura 4.1 suele denominarse *CCI empírica*.

Figura 4.1. CCI *empírica*. Probabilidad de acierto a un ítem en función de la puntuación tipificada (Z)



Desde la TRI se resume la información que contiene cada CCI empírica en una fórmula o modelo en el que (con uno, dos o tres parámetros del ítem) se recoge la información contenida en la función. Por tanto, en la aplicación de de la TRI, un paso inexcusable es optar por un modelo que sea una buena descripción del rendimiento en los ítems.

En la Figura 4.2 se representan dos de los muchos modelos que podrían aplicarse. En la figura superior se ha aplicado un modelo lineal que, en el ejemplo, sigue la siguiente ecuación:

$$P = 0{,}177Z + 0{,}492$$

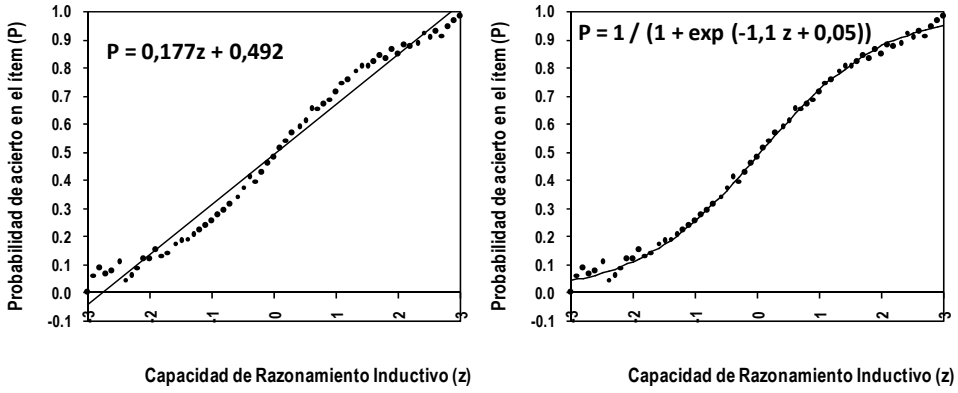
Un inconveniente de este modelo es que para niveles de rasgo extremos pueden obtenerse valores de P imposibles (negativos o mayores que uno).

En la figura inferior se ha aplicado un modelo logístico que, en el ejemplo, sigue la siguiente ecuación:

$$P = \frac{1}{1 + e^{-1{,}1Z + 0{,}05}}$$

donde e es la base de los logaritmos neperianos (2,718). En este modelo, el valor de P necesariamente estará comprendido entre 0 y 1. Esta es una de las razones por la que en TRI suelen aplicarse los modelos logísticos. Una de las características de los modelos logísticos es que la razón entre p y $1 - p$ se incrementa exponencialmente en relación a z. La forma exacta de la función exponencial dependerá de las características del ítem.

Figura 4.2. CCI según un modelo teórico. Modelo lineal (arriba) y Modelo logístico (abajo)



Modelos de TRI para ítems dicotómicos

Un problema importante es que la puntuación típica observada en el test, *Z,* puede no ser una buena medida del nivel de rasgo. Por ejemplo, si el test tiene un bajo coeficiente de fiabilidad; en ese caso, si se utilizara *Z* en el eje de abcisas, la CCI no representaría bien la relación entre el nivel de rasgo y la probabilidad de acertar el ítem. Por ello, en TRI se aplica el modelo utilizando el verdadero nivel de rasgo, al que se le denomina θ, que es una variable no observable (como lo era *V* en la TCT). Más adelante veremos cómo se pueden estimar las CCI siendo θ no observable. Pero antes debemos estudiar los distintos modelos logísticos que pueden dar cuenta de datos como los mostrados en la Figura 4.1.

Modelo logístico de un parámetro (ML1P)

Este es el modelo más simple de todos. Se asume que los ítems varían sólo en un parámetro de dificultad, al que se le denomina parámetro *b*. La expresión matemática es:

$$P_j(\theta) = \frac{1}{1 + e^{-Da(\theta - b_j)}} \qquad [4.1]$$

donde $P_j(\theta)$ es la expresión que utilizaremos a partir de ahora para referirnos a la probabilidad de que una persona acierte el ítem *j* en función de su nivel de rasgo θ. Así pues, los términos de la fórmula son:

- $P_j(\theta)$ Probabilidad de acertar el ítem *j* si el nivel de rasgo es θ.
- θ Nivel de rasgo o nivel de habilidad de la persona; cuanto mayor sea θ, manteniendo constantes los demás elementos de la fórmula, mayor será $P_j(\theta)$. Generalmente, se asume que θ está en una escala de puntuaciones típicas; por tanto, sus valores variarán generalmente entre –3 y 3.
- b_j Es el *parámetro de dificultad* del ítem *j*; a mayor valor b_j, manteniendo constantes los demás elementos de la fórmula, menor será $P_j(\theta)$. En el ML1P el valor de b_j indica el nivel de θ en el que la probabilidad de acertar el ítem es 0,5. Si el nivel de rasgo θ está en una escala de puntuaciones típicas, los valores de *b* variarán generalmente entre –3 y 3.
- *a* Parámetro de discriminación, que en este modelo se asume que es igual para todos los ítems (por ello, no aparece el subíndice *j*). Por tanto, en el ML1P el parámetro *a* es una constante e indica la mayor o menor inclinación o pendiente de la CCI cuando $\theta = b_j$. Esto significa que en el ML1P todos los ítems tienen la misma pendiente. Generalmente, si el nivel de rasgo θ está en una escala de puntuaciones típicas, puede tomar valores entre 0,3 y 2,5 (sólo uno de ellos para todos los ítems de un test) según los ítems sean más o menos discriminativos.
- *e* Base de los logaritmos neperianos (2,718).
- *D* Constante arbitraria (D = 1,702 ó 1)[1]. En lo que sigue, asumiremos que D = 1,702. Es importante que el investigador explicite siempre cual es el valor de *D*. Si se elige el valor D = 1,702, se dice que se está utilizando el modelo con *métrica normal.* Si se elige el valor D = 1, se dice que se está utilizando el modelo con *métrica logística.*

[1] El valor de *D* es arbitrario y no afecta al ajuste de la función. Lo habitual es elegir D = 1; sin embargo, algunos autores utilizan D = 1,702; cuando D = 1,702, la función logística, $f_L(z)$, es muy similar a otra función muy conocida, $F_N(z)$, la función de probabilidad acumulada de la distribución normal, Z ~ N(0 , 1), evaluada en *z:*

$$f_L(z) = \frac{1}{1 + e^{-1,702z}} \cong \frac{1}{\sqrt{2\pi}} \int_{Z=-\infty}^{Z=z} exp(-0,5Z^2)dZ = F_N(z)$$

Otra forma frecuente de presentar el ML1P es:

$$P_j(\theta) = \frac{1}{1 + e^{-(\theta - b_j)}} \qquad [4.2]$$

que elimina las constantes *a* y *D* del modelo. Ambas ecuaciones ([4.1] y [4.2]) son equivalentes (ver apéndice). El modelo expresado en la ecuación [4.2] suele denominarse *Modelo de Rasch.*

Ejemplo 4.1. Cálculo de la probabilidad de acierto en el ML1P

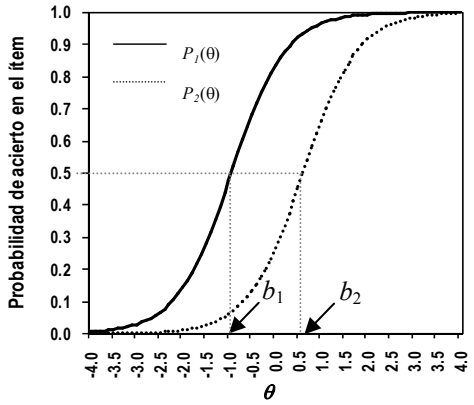
Una persona con nivel de habilidad $\theta = 1$ responde a un ítem *j* con parámetro de discriminación 1 y parámetro de dificultad 1 ($\theta = 1$, $a = 1$, $b_j = 1$), su probabilidad de acertar será:

$$P_j(\theta = 1) = \frac{1}{1 + e^{-1,702(1)(1-1)}} = \frac{1}{1 + e^{0}} = 0,5$$

El índice de dificultad (b_j) es, justamente, aquel valor de θ para el cual $P_j(\theta) = 0,5$. Por tanto, cuanto mayor sea *b* más difícil es el ítem.

En la Figura 4.3, se representan las CCI de dos ítems que difieren en dificultad.

Figura 4.3. CCI según el ML1P (*D* = 1,702; *a* = 1) para 2 ítems con $b_1 = -0,95$ y $b_2 = 0,6$.



En la primera, la que está más a la izquierda, el valor de θ al que corresponde $P_1(\theta) = 0,5$ es aproximadamente –0,95. Por lo tanto, la dificultad del primer item es $b_1 = -0.95$. En el

segundo ítem, el valor de θ al que corresponde $P_2(\theta) = 0{,}5$ es aproximadamente 0,6. Por lo tanto, la dificultad del segundo ítem es $b_2 = 0{,}6$. La Figura muestra que la probabilidad de acertar el ítem es sistemáticamente menor en el ítem 2 que en el ítem 1 para cualquier θ. El ítem 2 es más difícil que el uno, y sus índices de dificultad así lo muestran ($b_2 > b_1$).

Una interpretación de la probabilidad $P_j(\theta)$ es la siguiente: si $P_1(\theta = -0{,}95) = 0{,}5$ eso quiere decir que para una población con nivel de rasgo $\theta = -0{,}95$ el 50% acierta este ítem; o, también, que una persona de rasgo $\theta = -0{,}95$ acertará el 50% de los ítems con propiedades psicométricas iguales a las de este ítem.

En la Figura 4.3 puede observarse que las CCI de los dos ítems tienen la misma pendiente. Esta es una propiedad importante del ML1P: las CCI de distintos ítems nunca se cruzan por lo que el ordenamiento que hacemos de los ítems por su dificultad será siempre el mismo independientemente del grupo de personas con el que trabajemos; si un ítem es más fácil que otro, lo es para cualquier nivel de habilidad. De la misma manera, el ordenamiento que haremos de los evaluados por su nivel de habilidad será siempre el mismo, independientemente del conjunto de ítems que le apliquemos; si una persona, tiene más probabilidad de acertar un ítem que otra, también tendrá mayor probabilidad de acertar cualquier otro ítem, lo que hace mucho más clara la interpretación del significado de θ. Si se cumplen estas dos propiedades se habla de *Objetividad Específica* de las medidas. Veremos que en otros modelos de TRI no se cumplen.

Ejemplo 4.2. Modelo de Rasch

En la Tabla 4.1 se muestran los parámetros b de 3 ítems de una escala de cálculo numérico y los parámetros θ de 3 personas, obtenidos después de aplicar el modelo de Rasch. Una ventaja de la TRI frente a la TCT es que los parámetros θ de los personas y los parámetros b de los ítems están expresados en la misma escala; es decir, podemos comparar directamente el nivel θ de una persona con el parámetro b de un ítem. En el ejemplo, el evaluado 3 tiene un nivel de rasgo de 0. Su probabilidad de acertar los tres ítems puede deducirse de la relación entre su θ y el parámetro b de esos tres ítems; en concreto, su probabilidad de acertar el ítem 4 es alta (mayor que 0,5) ya que ese ítem tiene parámetro b menor que su θ; su probabilidad de acertar el ítem 3 es media (igual a 0,5) ya que ese ítem tiene parámetro b menor que su θ; su probabilidad de acertar el ítem 5 es baja (menor que 0,5) ya que ese ítem tiene mayor parámetro b menor que su θ. Tales comparaciones no son posibles desde la TCT, donde los índices de dificultad (p_j) y la puntuación en el test (X_j) están expresados en distinta escala.

Desde los modelos de Rasch, se facilita la interpretación de las puntuaciones de las personas. Si las operaciones cognitivas para resolver los ítems han sido bien delimitadas, podemos darle significado a cada nivel de rasgo en función de cuáles son las probabilidades de resolver exitosamente cada operación (implícita en cada ítem). Por ejemplo, el ítem 5 requiere que el estudiante sea capaz de resolver raíces cuadradas de una cierta complejidad. Un nivel de θ de 1,5 significa que existe una probabilidad de 0,5 de resolver ese tipo de raíces. Siguiendo ese razonamiento podemos llegar a una idea más exacta de qué competencias implica cada nivel de habilidad.

Tabla 4.1. Parámetros de tres evaluados y tres ítems según el modelo de Rasch aplicado para modelar las respuestas a una prueba de cálculo numérico.

θ del evaluado	*Evaluado*	*Valor*	*Ítem*	*b del ítem*	*Contenido del ítem*
		–2			
		–1,75	Ítem 4	–1,75	213,5 + 2,085 – 13,65 =
		–1,5			
		–1,25			
		–1			
-0,75	Sujeto 1	–0,75			
		–0,5			
		–0,25			
0	Sujeto 3	0	Ítem 3	0	2 (12 – 8) – 4 (2 – 4) =
		0,25			
0,5	Sujeto 2	0,5			
		0,75			
		1			
		1,5	Ítem 5	1,5	$\sqrt{157.2516}$ =
		2			

Modelo logístico de dos parámetros (ML2P)

Este modelo permite que el parámetro *a*, que indica la capacidad discriminativa del ítem, varíe de ítem a ítem:

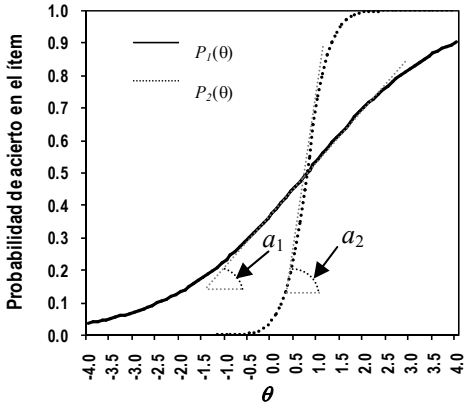
$$P_j(\theta) = \frac{1}{1 + e^{-Da_j(\theta - b_j)}} \qquad [4.3]$$

donde el parámetro a_j sigue siendo el *parámetro de discriminación*, pero en este modelo puede variar de ítem a ítem (por ello se añade el subíndice *j*). El parámetro a_j indica la mayor o menor inclinación o pendiente de la CCI cuando $\theta = b_j$. La pendiente en ese punto es exactamente $0{,}25Da_j$.

En la Figura 4.4 se han representado las CCI de dos ítems de igual dificultad ($b_1 = b_2 = 0{,}75$), pero que difieren en el parámetro de discriminación. El parámetro *a* se relaciona con la pendiente; es decir, es proporcional al ángulo que forma la CCI en relación al eje de abcisas. La principal diferencia entre los dos ítems es que el 2 (línea de puntos), cuando $\theta = 0{,}75$, tiene una pendiente mucho mayor ($a_2 = 2{,}4$) que la del ítem 1 ($a_1 = 0{,}4$). Como la pendiente es tan alta, las personas con $\theta > 0{,}75$ tienen casi todas ellas una muy alta probabilidad de acertar el ítem 2 (y casi todas ellas lo acertarán), y las personas con $\theta < 0{,}75$ tienen casi todas ellas una probabilidad próxima a cero de acertarlo (y casi ninguna lo acertará). Por lo tanto, el ítem 2 discrimina entre los que tienen $\theta > 0{,}75$ y los que tienen $\theta < 0{,}75$. Por su parte, el ítem 1 tiene muy poca pendiente cuando $\theta = 0{,}75$. En consecuen-

cia, aunque la mayoría de las personas con $\theta > 0{,}75$ lo acertarán, muchas lo fallarán (pues la probabilidad de acierto es claramente inferior a uno). Igualmente, aunque la mayoría de las personas con $\theta < 0{,}75$ fallarán el ítem, muchas lo acertarán, pues la probabilidad de acierto es claramente superior a cero. En el ítem 1 la probabilidad crece muy suavemente a medida que aumenta θ por lo que no es buen discriminador entre las personas con $\theta > 0{,}75$ y las que tienen $\theta < 0{,}75$.

Figura 4.4. CCI según el ML2P para 2 ítems ($a_1 = 0{,}4$; $b_1 = 0{,}75$; $a_2 = 2{,}4$; $b_2 = 0{,}75$).



Los valores de a oscilaran generalmente entre 0,3 y 2,5, y se suelen considerar items discriminativos los que tienen valores a mayores de uno. El parámetro b_j se interpreta, en este modelo, de la misma manera que en el ML1P.

Modelo logístico de tres parámetros (ML3P)

Este modelo añade a los dos parámetros a y b un tercero, c, que representa la probabilidad de acertar el ítem al azar. Más exactamente, c es el valor de $P_j(\theta)$ para valores extremadamente bajos de θ. La expresión del modelo de 3 parámetros es la siguiente:

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1 + e^{-Da_j(\theta - b_j)}} \qquad [4.4]$$

Los parámetros en la ecuación [4.4] se interpretan en este caso de la siguiente manera:

1. El **parámetro c_j de pseudoazar** representa la probabilidad de acierto para personas con un nivel de rasgo extremadamente bajo. Si no hay omisiones, suele tomar un valor próximo al inverso del número de opciones de respuesta (algo menor si se descartan opciones incorrectas con facilidad). Su valor también depende de la presencia de omisiones: cuanto mayor sea el número de personas que no responden al ítem, menor será el parámetro c. Como es una probabilidad, sus valores pueden oscilar entre 0 y 1, pero generalmente varían entre 0 y 0,5.

2. El **parámetro b_j de dificultad**, indica la posición de la CCI en relación al eje de abcisas (cuanto mayor la dificultad del ítem, más hacia la derecha se posiciona la CCI). Se encuentra en la misma métrica que θ, por lo que sus valores suelen oscilar en el mismo rango. Indica el nivel de habilidad θ donde la probabilidad de acertar es el valor medio entre c_j y 1; es decir, $0{,}5(1 + c_j)$. Este es el punto de máxima discriminación del ítem (es decir el punto donde la pendiente de la CCI es máxima).
3. El **parámetro a_j de discriminación** es proporcional a la pendiente que tiene la CCI en el valor intermedio $\theta = b_j$.[2] Los valores de a suelen oscilar entre 0,3 y 2,5 (según la métrica del nivel de rasgo que hemos fijado).

Debemos observar que los parámetros de dificultad y discriminación no son iguales a los del modelo ML2P[3].
En la Figura 4.5 podemos ver la CCI de dos ítems con los mismos valores de a (1) y b (0), pero distintos valores de parámetro c (c_1= 0 y c_2= 0,2).

[2] En concreto, la pendiente en el punto $\theta = b_j$ depende de a_j y de c_j y es $0{,}25Da_j(1 - c_j)$.

[3] Definamos que la probabilidad de acertar como función del nivel de rasgo *si no hubiera aciertos por azar* sigue el ML2P:

$$P_j^*(\theta)=\frac{1}{1+e^{-Da_j(\theta-b_j)}}$$

En condiciones donde hay respuestas al azar el ML2P es inadecuado. Pues bien, el modelo de 3 parámetros incluye al anterior modelo un nuevo parámetro c_j que indica la probabilidad de aciertos por azar:

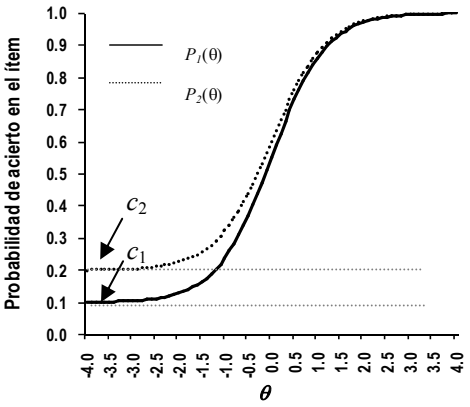
$$P_j(\theta) = P_j^*(\theta) + \left(1 - P_j^*(\theta)\right)c_j$$

La lógica del modelo de la ecuación es la siguiente. La probabilidad de acertar es la probabilidad de saber la respuesta [$P^*_j(\theta)$], más la probabilidad de no saberla [$1 - P^*_j(\theta)$] multiplicada por la probabilidad de acertarla cuando no se sabe la respuesta [c_j]; el parámetro c_j sirve para modelar el hecho de que aunque uno no sepa la respuesta, puede acertarla eligiendo al azar entre las opciones o escogiendo aquella que resulte más atractiva. La expresión se simplifica a la ecuación [4.4]:

$$P_j(\theta) = P_j^*(\theta) + (1 - P_j^*(\theta))c_j = c_j + (1 - c_j)P_j^*(\theta) = c_j + (1 - c_j)\frac{1}{1+e^{-Da_j(\theta-b_j)}}$$

Observe que aunque el desarrollo del ML3P se inspira en el ML2P, las ecuaciones de ambos modelos son distintas, [4.3] y [4.4], por lo que los parámetros a_j y b_j serán también distintos.

Figura 4.5. CCI según el ML3P para 2 ítems ($a_1 = 1$; $b_1 = 0$; $c_1 = 0$; $a_2 = 1$; $b_2 = 0$; $c_2 = 0{,}2$)



Desde nuestro punto de vista, el modelo más completo es el ML3P. En el modelo de Rasch y en el ML2P no se contemplan las respuestas al azar. Esto hace que ambos puedan ajustar peor cuando se trabaja con ítems de opción múltiple, especialmente para ítems de dificultad elevada. Además, en el modelo de Rasch, tampoco se contempla la posibilidad de que los ítems tengan distinto parámetro de discriminación. Esto supone asumir que no hay ítems peores y mejores para medir el nivel de rasgo (o, en todo caso, que aquellos ítems cuyo parámetro *a* se diferencie mucho del de los otros ítems deberían ser eliminados de la prueba). El modelo de Rasch raramente ajusta a los datos si no es mediante una criba de ítems que, finalmente, puede acabar por socavar la validez del test. Por otro lado, la aplicación del ML3P requiere procedimientos más complejos de estimación de los parámetros y muestras más numerosas. Los que defienden el modelo de Rasch se basan en algunas de sus recomendables propiedades estadísticas (p.ej., la objetividad específica o, como veremos más adelante, que proporciona estimadores suficientes de los parámetros[4]). Además, justifican que si un modelo más parsimonioso (con menos parámetros) se ajusta a los datos, es preferible a modelos más complejos.

A partir de una CCI conoceremos también la probabilidad de fallar el ítem, a la que nos referiremos como $Q_j(\theta)$. Más genéricamente, podremos referirnos a la probabilidad de una respuesta x_j al ítem j lo que suele expresarse como:

$$P_j(X_j = x_j \mid \theta) = P_j(\theta)^{x_j} Q_j(\theta)^{1-x_j} \qquad [4.5]$$

que es una forma compacta de referirse a la probabilidad de la respuesta x_j; observe que la formula anterior se simplifica en cada caso al resultado correcto:

$$P_j(X_j = 1 \mid \theta) = P_j(\theta)^1 Q_j(\theta)^0 = P_j(\theta)$$
$$P_j(X_j = 0 \mid \theta) = P_j(\theta)^0 Q_j(\theta)^1 = Q_j(\theta)$$

[4] Se dice de un estimador que es suficiente si agota toda la información disponible en la muestra para estimar el parámetro.

Supuestos de la TRI

Un paso previo a la aplicación de los modelos de TRI es la comprobación de que se cumplen sus dos supuestos fundamentales: unidimensionalidad e independencia local. En este apartado se describen los dos supuestos y por qué son importantes. En los capítulos 6, 10 y 11 se profundizará en los procedimientos para comprobar ambos supuestos.

Unidimensionalidad

En los modelos anteriores la probabilidad de acertar un ítem depende únicamente de sus parámetros y de θ. Por ejemplo, en un ítem que mida el nivel de vocabulario inglés, la probabilidad de acertarlo depende de los valores *a*, *b* y *c* del ítem y del nivel de vocabulario en inglés de la persona (θ), pero no de otros rasgos, como podría ser su inteligencia. En otras palabras, se asume que el rendimiento en los ítems que forman el test depende del nivel de la persona en un solo rasgo o dimensión. A este supuesto se le denomina *supuesto de unidimensionalidad*. La mayoría de las definiciones actuales de unidimensionalidad hacen referencia al análisis factorial y al concepto de *independencia local débil:*

$$\sigma_{X_j X_{j'}|\theta} = 0 \qquad [4.6]$$

que implica que las covarianzas entre ítems para muestras con el mismo nivel de rasgo son cero. En otras palabras, cumpliéndose el supuesto, si seleccionáramos a un grupo de evaluados con el mismo nivel de rasgo la correlación entre dos ítems cualesquiera sería cero. Según los modelos, dos ítems correlacionan sólo porque acertarlos depende de θ; por tanto, si condicionamos los datos en dos ítems a los valores θ debe desaparecer la correlación. En los Capítulos 6 y 10 se estudiarán los procedimientos de análisis factorial que permiten estudiar si se cumple el supuesto de unidimensionalidad.

Independencia local

Existe *independencia local* entre los ítems de un test si la respuesta de una persona a uno de ellos no depende de sus respuestas a los otros. La independencia local se deriva de la unidimensionalidad porque significa que la respuesta a un ítem sólo depende de sus parámetros y de θ, y no está influida por el orden de presentación de los ítems, las respuestas que ya se hayan dado, etc. Para modelos como los descritos, la unidimensionalidad implica independencia local; sin embargo, conviene mantener separados ambos supuestos, ya que en los modelos multidimensionales de TRI no son equivalentes. Matemáticamente, la independencia local se define en términos probabilísticos: la probabilidad de que un evaluado *i* tenga un patrón de respuestas en un test de *J* ítems es igual al producto de las probabilidades de cada respuesta en cada uno de ellos por separado:

$$P(X_1 = x_{i1}, X_2 = x_{i2}, ...X_k = x_{ik} ...X_J = x_{iJ} \mid \theta) = \prod_{j=1}^{J} P(X_j = x_{ij} \mid \theta)$$

o, de forma más compacta:

$$P(\mathbf{X_i}) = \prod_{j=1}^{J} P(X_j = x_{ij} \mid \theta) \qquad [4.7]$$

donde $P(\mathbf{X_i}|\theta)$ designa la probabilidad del patrón de respuestas para el evaluado *i*; $\mathbf{X_i}$ se define como un vector con las respuestas del evaluado *i,* $\mathbf{X_i} = \{x_{i1}, x_{i2}, x_{i3} \ldots x_{iJ}\}$[5]. Este planteamiento se conoce también como supuesto de *independencia local fuerte.* Gracias a este supuesto se cumple que, conociendo los parámetros del evaluado y de los ítems, podemos hallar la probabilidad de su patrón de respuestas en el test completo.

El supuesto de independencia local fuerte, como vemos, está referido a todos los ítems del test y por tanto es difícil de contrastar empíricamente. Por ello se suele contrastar el supuesto de independencia local débil*,* ya mencionado, relativo sólo a pares de ítems. Tal supuesto implica que, para cualquier par de ítems, se cumple que las probabilidades de respuesta son independientes para evaluados con el mismo nivel de rasgo θ:

$$P(X_1 = x_{i1}, X_2 = x_{i2} \mid \theta) = P_1(X_1 = x_{i1} \mid \theta) P_2(X_2 = x_{i2} \mid \theta) \qquad [4.8]$$

que también puede expresarse, en el caso de ítems dicotómicos, como (ver ecuación [4.5]):

$$P(X_1 = x_{i1}, X_2 = x_{i2} \mid \theta) = P_1(\theta)^{x_{i1}} Q_1(\theta)^{1-x_{i1}} P_2(\theta)^{x_{i2}} Q_2(\theta)^{1-x_{i2}}$$

En el caso de ítems dicotómicos, es fácil observar que si se cumple lo anterior se cumple que la covarianza entre los ítems es 0 para evaluados con el mismo nivel de rasgo. En efecto, la covarianza entre los ítems 1 y 2 se calcula como:

$$\sigma_{X_1 X_2 | \theta} = P(X_1 = 1, X_2 = 1 \mid \theta) - P(X_1 = 1 \mid \theta) P(X_2 = 1 \mid \theta)$$

Si hay independencia local los dos términos a la derecha de la ecuación son iguales. Por el contrario, si hay *dependencia local*:

$$\sigma_{X_j X_{j'} | \theta} \neq 0$$

La dependencia local puede ser positiva o negativa. Si es positiva ($\sigma_{X_j X_{j'} | \theta} > 0$), el número de personas con la misma respuesta en los dos ítems es mayor que el esperado según el modelo unidimensional. Generalmente, ítems con dependencia local positiva miden una

[5] Las variables en negrita se utilizan para designar una matriz o un vector.

misma dimensión específica distinta de θ. Por ejemplo, puede ocurrir que dos ítems tengan un enunciado similar y sean redundantes o que para su resolución requieran una destreza que no requieren otros ítems del test; si se aplica un modelo de TRI a estos datos, los ítems parecerán más discriminativos de lo que realmente son y se sobrestimará la precisión de la prueba. Si la dependencia local es negativa ($\sigma_{X_j X_{j'}|\theta} < 0$), esto quiere decir que cuando una persona tiende a rendir mejor de lo esperado en un ítem, tiende a rendir peor en otro ítem (y viceversa). Generalmente, ítems con dependencia local negativa miden dimensiones distintas.

Ejemplo 4.3. Concepto de Independencia Local

Un test consta de dos ítems y la probabilidad de que un evaluado J acierte el primero es $P_1(\theta) = 0{,}4$ y la de que acierte el segundo $P_2(\theta) = 0{,}8$. El principio de independencia local establece que la probabilidad de que acierte los dos viene dada por:

$$P_1(\theta)P_2(\theta) = (0{,}4)(0{,}8) = 0{,}32$$

La probabilidad de acertar el primero y fallar el segundo sería:

$$P_1(\theta)Q_2(\theta) = (0{,}4)(0{,}2) = 0{,}08$$

La de que falle el primero y acierte el segundo será:

$$Q_1(\theta)P_2(\theta) = (0{,}6)(0{,}8) = 0{,}48$$

La de que falle ambos ítems será:

$$Q_1(\theta)Q_2(\theta) = (0{,}6)(0{,}2) = 0{,}12$$

Supongamos que 100 personas con idéntico nivel de rasgo que la persona J, $\theta = 0$, contestan al test. Se esperarían aproximadamente los resultados de la Tabla 4.2.

Tabla 4.2. Número de personas con cada patrón de respuestas (1, acierto; 0, error) si se cumple la independencia local

Ítem 1	*Ítem 2*	*Nº personas esperado si se cumpliera la independencia local*
1	1	(0,4)(0,8)(100) = 32
1	0	(0,4)(0,2)(100) = 8
0	1	(0,6)(0,8)(100) = 48
0	0	(0,6)(0,2)(100) = 12

Si correlacionamos las 100 respuestas al primer ítem con las 100 respuestas al segundo, el resultado sería cero, lo que indicaría que se cumple el supuesto de independencia local; es decir, que $\sigma_{X_jX_j|\theta} = 0$):

$$\sigma_{X_1X_2|\theta=0} = P(X_1 = 1, X_2 = 1 \mid \theta = 0) - P(X_1 = 1 \mid \theta = 0)P(X_2 = 1 \mid \theta = 0) =$$
$$= \frac{32}{100} - \frac{40}{100}\frac{80}{100} = 0$$

Estimación de parámetros

En la práctica, una vez que se han comprobado los supuestos de unidimensionalidad e independencia local, el siguiente paso es aplicar un modelo de TRI, lo que requiere un método estadístico para estimar los parámetros de los evaluados y de los ítems. Seleccionado un modelo de TRI, hay que aplicar el test a una muestra amplia y, a partir de la matriz de respuestas obtenidas, estimar los parámetros de cada ítem y la θ de cada evaluado. La estimación de parámetros es el paso que nos permite llegar desde las respuestas conocidas de las personas a los ítems hasta los valores desconocidos de los parámetros de los ítems y de los niveles de rasgo.

El concepto de estimación máximo verosímil (ML)

Para obtener las estimaciones se aplica fundamentalmente el método de máxima verosimilitud (ML)[6], mediante el cual se encuentran los valores de los parámetros que hagan más probable la matriz de respuestas obtenida. La estimación de los parámetros en TRI supone un proceso complejo. La mejor referencia para una descripción detallada de todos los pro-

[6] Veamos a continuación un ejemplo sencillo de estimación ML. Si lanzamos una moneda diez veces y obtenemos siete caras, el estimador ML del parámetro *p* (probabilidad de cara de la moneda) es 7/10 = 0,7. El resultado "siete caras en diez lanzamientos" es poco compatible con que la probabilidad de cara sea 0,1, o 0,2. De hecho, la probabilidad de obtener siete caras y tres cruces es prácticamente cero si *p* = 0,1 o si *p* = 0,2. Dicha probabilidad pasa a ser 0,117 si *p* = 0,5, y alcanza el máximo valor (0,267) cuando *p* = 0,7. El estimador ML proporciona el valor de *p* bajo el que es máxima la probabilidad del suceso que se ha observadoLa probabilidad de *x* caras en *n* lanzamientos sigue la distribución binomial:

$$B(x;n,p) = \binom{n}{x} p^x q^{1-x}$$

donde el primer término después de la igualdad, el número combinatorio, indica el número de formas en las que pueden surgir *x* caras en *n* lanzamientos. En el ejemplo, hay 120 maneras distintas de obtener 7 caras en 10 lanzamientos. Si la probabilidad de cara es 0,7 entonces la probabilidad de que se obtengan 7 caras en 10 lanzamientos es 0,267:

$$B(x = 7; n = 10, p = 0{,}7) = \frac{10!}{7!(10-3)!} 0.7^7 0.3^3 = 120(0{,}00222) = 0{,}267$$

cedimientos de estimación la encontramos en los libros de Baker (p.ej., Baker y Kim, 2004). Información similar en castellano puede encontrarse en López- Pina (1995) o también en Revuelta, Abad y Ponsoda (2006).

Fases en el proceso de estimación de los parámetros

En TRI, se pueden distinguir dos objetivos de estimación:

1. La primera vez que se aplica un test hay que estimar conjuntamente los parámetros de los ítems y los parámetros θ. Ese proceso se conoce como *calibración de los ítems*. La fase de calibración es la de mayor complicación puesto que hay que hacer asunciones sobre la distribución del nivel de rasgo y se requieren muestras numerosas. Si tenemos, por ejemplo, diez ítems que miden un mismo rasgo, los podemos aplicar a una muestra de 300 personas. La matriz de datos tendrá 300 filas, siendo cada fila la secuencia de unos (aciertos) y ceros (errores) de cada persona de la muestra. Si queremos aplicar el ML3P, tendremos que estimar los 30 parámetros de los ítems (es decir, a, b y c de cada ítem) y 300 parámetros de las personas (los 300 valores θ, uno por persona).
2. Una vez que son estimados los parámetros de los ítems, pueden considerarse conocidos y usados en posteriores aplicaciones para estimar el nivel de rasgo de las personas. Hablaremos entonces de *estimación del nivel de rasgo*.

A continuación, se ilustra cada una de las fases. Empezaremos por el caso más simple, la estimación del nivel de rasgo.

Estimación del nivel de rasgo por el método ML

En TRI, el procedimiento de estimación sigue una lógica similar al comentado para la moneda. Supongamos, por ejemplo, que tenemos un test compuesto por tan sólo dos ítems para los que ya conocemos sus parámetros ($b_1 = –0{,}7$; $b_2 = 1$), y que lo aplicamos a una persona. Supongamos también que acierta el primero y falla el segundo. A partir de estas respuestas la estimación ML de su θ se puede explicar de forma gráfica. Como el evaluado ha acertado el primer ítem, podemos calcular, mediante su CCI (recuérdese que los parámetros del ítem son conocidos), la probabilidad de que esto ocurra para cada nivel de θ. Esto se muestra en la Figura 4.6.

Como el evaluado ha fallado el segundo ítem, a partir de su CCI podemos calcular la probabilidad de que esto ocurra para cada uno de los valores de θ. En concreto, como la probabilidad de fallar, $Q_2(\theta)$, se puede obtener a partir de la probabilidad de acertar, podremos representar la probabilidad de error en el segundo ítem como se muestra en la Figura 4.7. Nótese que no se representa la CCI del ítem 2, pues para cada valor de θ se ha representado la probabilidad de error y no la de acierto. Puede observarse que es más probable que fallen el ítem los evaluados con niveles bajos de habilidad que los evaluados con niveles altos (cosa bastante lógica).

Figura 4.6. Probabilidad de acertar el ítem 1 con parámetro $b_1 = -0,7$

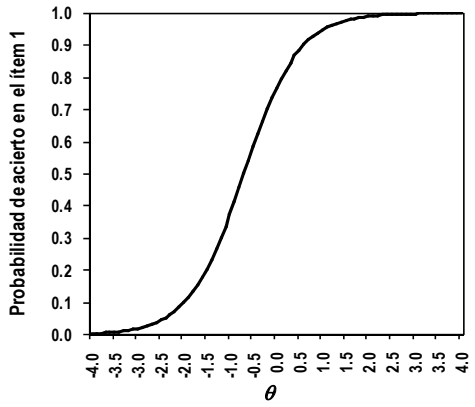
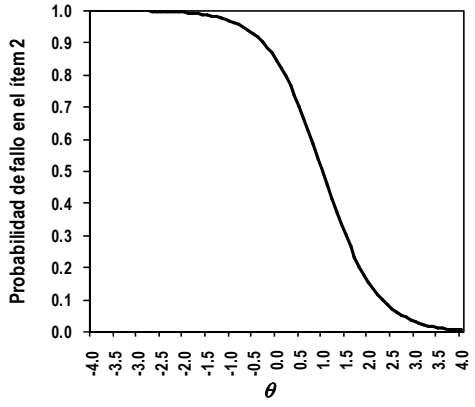


Figura 4.7. Probabilidad de fallar el ítem 2 con parámetro $b_2 = 1$



El valor estimado de θ para esta persona sería aquel que haga más probable el resultado obtenido (acertar el primer ítem y fallar el segundo). Según el supuesto de independencia local, ambos sucesos son independientes y, por lo tanto, la probabilidad de que ocurran ambos conjuntamente es igual al producto de las probabilidades de acertar el primero, $P_1(\theta)$, por la de fallar el segundo, $Q_2(\theta)$ (ver ecuación [4.8]).

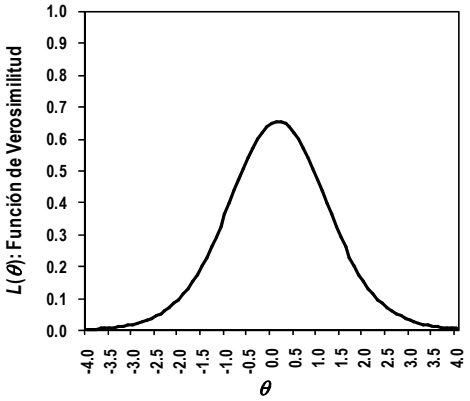
La probabilidad del patrón de respuestas se denomina en este contexto *función de verosimilitud* (para el evaluado i) y se designa como $L_i(\theta)$, que indica la probabilidad de las respuestas de un evaluado dado un valor del parámetro θ, siendo conocidos los parámetros a, b y c de los ítems. En nuestro caso:

$$L_i(\theta) = P_1(\theta)Q_2(\theta)$$

Si representamos gráficamente la función $L(\theta)$ para cada valor de θ, obtendríamos la Figura 4.8. En este caso vemos que el valor θ que hace más probable el resultado obtenido

(acierto en el primer ítem y fallo en el segundo) es algo mayor que cero. De hecho, 0,15 será la θ estimada para esta persona.

Figura 4.8. Probabilidad de acertar el ítem 1 y fallar el ítem 2



En general, se responderá a un número de ítems mayor de dos y se producirán particulares secuencias de unos y ceros. La probabilidad de obtener tal secuencia de aciertos y errores para un evaluado *i* se puede expresar como:

$$L_i(\theta) \equiv P(\mathbf{X}_i \mid \theta) = \prod_{j=1}^{J} P_j(\theta)^{x_{ij}} Q_j(\theta)^{1-x_{ij}} \qquad [4.9]$$

El nivel de rasgo estimado por el *método de máxima verosimilitud (ML)* será el valor θ para el que la anterior expresión alcanza su máximo valor.

Ejemplo 4.4. Estimación del nivel de rasgo por el método ML

Un test consta de 4 ítems, cuyos parámetros, según el modelo de Rasch, son –1, 0, 1 y 2. Una persona completa el test y acierta los tres primeros ítems y falla el cuarto. Puede obtenerse el valor de la función de verosimilitud, $L_i(\theta)$, para los siguientes valores θ: –3, –2, –1, 0, 1, 2 y 3, y así comprobar cuál de ellos maximiza $L_i(\theta)$. Aplicando la fórmula del ML1P se obtiene la probabilidad de acierto para cada ítem y cada uno de los valores de θ (ver Tabla 4.3). La función de verosimilitud, $L_i(\theta)$, al haber acertado los 3 primeros ítems y fallado el último, será la siguiente:

$$L_i(\theta) = P_1(\theta)P_2(\theta)P_3(\theta)Q_4(\theta)$$

Al aplicar la fórmula se obtiene $L_i(\theta)$ para cada valor de θ. Por ejemplo, para $\theta = 2$:

$$L_i(\theta = 2) = (0{,}99)(0{,}97)(0{,}85)(0{,}50) = 0{,}41$$

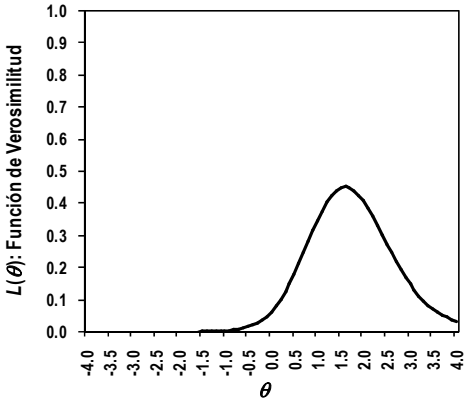
En la última fila de la Tabla 4.3 se muestra el valor de $L_i(\theta)$ para cada valor de θ.

Tabla 4.3. Probabilidad de la respuesta dada al ítem como función del nivel de θ

Ítems	b	Respuesta	θ	−3	−2	−1	0	1	2	3
1	−1	1	$P_1(\theta)$	0,03	0,15	0,50	0,85	0,97	0,99	1,00
2	0	1	$P_2(\theta)$	0,01	0,03	0,15	0,50	0,85	0,97	0,99
3	1	1	$P_3(\theta)$	0,00	0,01	0,03	0,15	0,50	0,85	0,97
4	2	0	$Q_4(\theta)$	1,00	1,00	0,99	0,97	0,85	0,50	0,15
			$L_i(\theta)$	0,00	0,00	0,00	0,06	0,35	0,41	0,15

Por lo tanto, de los siete valores θ, el que maximiza $L_i(\theta)$ es $\theta = 2$. En realidad se trata de una aproximación pobre, porque sólo se ha hecho la comprobación para 7 valores de rasgo. Cuando se trata de estimar en una situación real el nivel de rasgo, no se hace una búsqueda restringida a unos cuantos valores. En la Figura 4.9 se muestran los valores $L_i(\theta)$ para todas los valores θ comprendidos entre –4 y 4.

Figura 4.9. Probabilidad de acertar los tres primeros ítems y fallar el cuarto



En este ejemplo, el valor θ que maximiza $L_i(\theta)$ es 1,6 (ver Figura 4.9). Por tanto, la puntuación estimada para esta persona sería 1,6.

En la TRI, se obtiene el máximo de $L_i(\theta)$ por métodos numéricos, mediante programas de ordenador que contienen algoritmos que encuentran el valor θ para el que la función $L_i(\theta)$ alcanza el valor máximo. Para ello, se utiliza otra función que tiene el mismo máximo, *Ln* $L_i(\theta)$, más tratable matemáticamente:

$$LnL_i(\theta) = \sum_{j=1}^{J} \left[x_{ij} LnP_j(\theta) + (1 - x_{ij}) LnQ_j(\theta) \right] \quad [4.10]$$

Para obtener el máximo de una función puede calcularse la derivada de esa función (recuerde que si la derivada de una función en un punto es cero, la función tiene un máximo, un mínimo o un punto de inflexión). Se busca el parámetro θ para el que se satisface la ecuación:

$$\frac{\delta}{\delta\theta} LnL_i(\theta) = 0 \quad [4.11]$$

La derivada de $Ln\ L_i(\theta)$, en el caso del ML3P, es:

$$\frac{\delta}{\delta\theta} LnL_i(\theta) = D \sum_{j=1}^{J} a_j \frac{P_j^*(\theta)}{P_j(\theta)} (x_{ij} - P_j(\theta)) \quad [4.12]$$

donde $P^*_j(\theta)$ se define como:

$$P_j^*(\theta) = \frac{1}{1 + e^{-Da_j(\theta - b_j)}} \quad [4.13]$$

y donde a_j y b_j son los parámetros de discriminación y dificultad estimados en el ML3P. El máximo en $Ln\ L_i(\theta)$ se obtiene para el valor de θ en el que la derivada es cero. Esto ocurre cuando la suma ponderada de las diferencias $[x_{ij} - P_j(\theta)]$ se aproxima a 0 (ver ecuación [4.12]). La ponderación refleja que se da más importancia a los ítems más discriminativos y a aquellos en los que la diferencia entre $P^*_j(\theta)$ y $P_j(\theta)$ es más pequeña, lo que ocurre cuando ambas probabilidades son altas (el término $P^*_j(\theta)/P_j(\theta)$ oscilará entre 0, para niveles muy bajos y 1 para niveles muy altos de rasgo).

Para el ML2P se tendría que:

$$\frac{\delta}{\delta\theta} Ln(L_i(\theta)) = D \sum_{j=1}^{J} a_j (x_{ij} - P_j(\theta)) \quad [4.14]$$

Indicando que se ponderan más las respuestas a los ítems más discriminativos. Mientras que para el ML1P, tendríamos:

$$\frac{\delta}{\delta\theta} Ln(L_i(\theta)) = Da \sum_{j=1}^{J} (x_{ij} - P_j(\theta)) \quad [4.15]$$

Observe que en el modelo de Rasch, la ecuación [4.15] se simplifica a:

$$\frac{\delta}{\delta\theta} Ln(L_i(\theta)) = \sum_{j=1}^{J} (x_{ij} - P_j(\theta)) \quad [4.16]$$

Por tanto, en este último caso todos los ítems tendrían la misma importancia para la estimación; el valor θ estimado será aquel que haga que el número esperado de aciertos coincida con el número observado. Es decir, aquella θ para la que se cumple la igualdad:

$$\sum_{j=1}^{J} x_{ij} = \sum_{j=1}^{J} P_j(\theta) \quad [4.17]$$

Y, en definitiva, puesto que la suma de las puntuaciones en los ítems es la puntuación en el test, será aquella θ para la que se cumple la igualdad:

$$X_i = \sum_{j=1}^{J} P_j(\theta) \quad [4.18]$$

Calibración de los ítems

Partiendo de que tanto los parámetros de los ítems como los parámetros de los evaluados son desconocidos, existen distintos procedimientos para estimar los parámetros de los ítems. De nuevo, se trata de estimar los parámetros *a*, *b* y *c* de los ítems que maximizan la probabilidad de las respuestas observadas. Para ello, es necesario el uso de programas informáticos específicos. En el capítulo 11, se describen los procedimientos y algunos de los programas disponibles para la estimación de parámetros en TRI.

Ejemplo 4.5. Calibración de los ítems de un test de cálculo numérico

Una prueba de siete ítems de cálculo numérico (Tabla 4.4) ha sido respondida por 2.000 estudiantes. Al aplicar la TRI obtenemos las estimaciones de los parámetros de los ítems que se muestran en la Tabla 4.5. Puede observarse que el parámetro *a* guarda una relación directa con la correlación ítem-test de la TCT (r = 0,75), mientras que el parámetro *b* guarda una relación inversa con el índice de dificultad clásico o proporción de aciertos (r = –0,97). El ítem más fácil es el primero (menor parámetro *b*) y más difícil el quinto (mayor parámetro *b*). Los ítems más discriminativos son el 6 y el 7 (que se refieren a series numéricas). El ítem 7 tiene el mayor parámetro *c*; podría ser que en este ítem la opción correcta *d)* es atractiva para quien no sabe la respuesta (dado que –42 es el número más próximo a –40). Por otro lado, dados los parámetros de los ítems, concluiríamos que aplicar el ML1P sería inadecuado pues sólo los ítems 3, 4 y 5 tienen un parámetro *c* bajo y un parámetro *a* similar. En la Figura 4.10, se muestran las CCI de los 7 ítems.

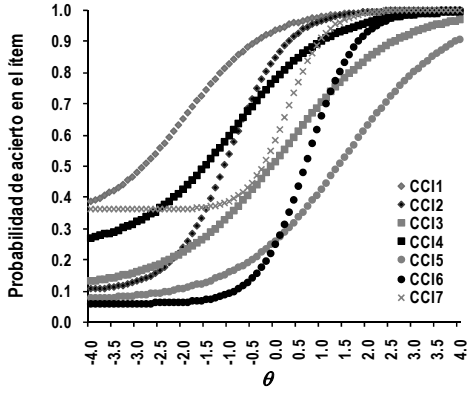
Tabla 4.4. Siete ítems de una prueba de cálculo numérico

Ítem	*Opciones*			
	a)	*b)*	*c)*	*d)*
1. ¿Cuál es el resultado de la siguiente operación? 2 + 8 – 15 + 9 – 7 – 3	−16	6	**−6**	−26
2. ¿Cuál es el resultado de la siguiente operación? (125/5) – (2)(2,5) + 2,455	**2,475**	−2,425	2,425	2,485
3. ¿Cuál es el resultado de la siguiente operación? 2 (12 – 8) – 4 (2 – 4))	**16**	0	24	−8
4. ¿Cuál es el resultado de la siguiente operación? 213,5 + 2,085 – 13,65	204,235	203,065	202,035	**201,935**
5. La raíz cuadrada de 157,2516 es:	12,455	12,554	12,45	**12,54**
6. Siga la serie 12,3, 14, 15,7, 17,4,… hasta encontrar el término que (por defecto o por exceso) se aproxime más a 22 ¿Cuál es el término?	21,5	**22,5**	20,8	22,4
7. Siga la serie −78, −69, -60,… hasta encontrar el término que (por defecto o por exceso) se aproxime más a −40 ¿Cuál es el término?	−52	−51	−33	**−42**

Tabla 4.5. Parámetros según la TCT y la TRI con el modelo logístico de 3 parámetros (métrica normal)

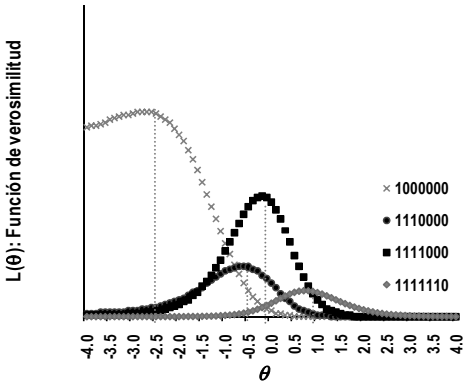
	P	r^{c}_{bp}	*a*	*b*	*c*
1	0,91	0,20	0,68	−1,92	0,33
2	0,78	0,35	0,97	−0,97	0,10
3	0,52	0,24	0,51	0,20	0,11
4	0,76	0,24	0,59	−0,88	0,24
5	0,29	0,21	0,52	1,53	0,07
6	0,34	0,34	1,14	0,71	0,06
7	0,62	0,30	1,42	0,26	0,36

Figura 4.10. CCI según el modelo logístico de tres parámetros



En definitiva, la TRI nos permite reproducir mediante un modelo cuál va a ser la proporción de aciertos en cada nivel de rasgo, algo que no proporcionaba la TCT. Además, puede estimarse θ en función del patrón de respuestas a los ítems. En la Figura 4.11 se han representado las funciones de verosimilitud asociadas a 4 patrones de respuesta.

Figura 4.11. Máximo de la función de verosimilud, $L_i(\theta)$ para cuatro patrones de respuestas



Los valores θ estimados por máxima verosimilitud para cada uno de esos patrones de respuesta serían, respectivamente –2,740 ("fallar los seis últimos ítems"), –0,633 ("fallar los cuatro últimos ítems"), –0,205 ("fallar los tres últimos ítems") y 0,740 ("fallar el último ítem").

Bondad de ajuste: Comparación de las CCI teóricas y las CCI observadas

La mayor parte de los programas informáticos de TRI incluyen estadísticos y residuos que permiten cuantificar la discrepancia entre los datos observados y los esperados si el modelo fuera correcto. Un modelo de TRI sólo puede aplicarse a unos datos, si estos datos se ajustan al modelo. La estrategia más utilizada para estudiar el ajuste es, para cada ítem, obtener el grado de discrepancia entre las probabilidades teóricas y empíricas de escoger cada opción de respuesta, condicionadas al nivel de rasgo. Tradicionalmente, se ha propuesto agrupar a las personas en Q intervalos según su nivel de rasgo estimado (p.ej., 10 intervalos). La agrupación se hace de forma que en cada intervalo haya un número mínimo de personas (p.ej., 5). En este caso, la proporción observada de aciertos en cada intervalo (O_q) se obtendría simplemente como la proporción observada de aciertos en el grupo q; la probabilidad teórica (E_q) es la probabilidad de acierto que predice el modelo, según la curva característica del ítem, para la media o la mediana del nivel de rasgo estimado en ese intervalo. Posteriormente, para cada ítem se obtiene un estadístico G^2:

$$G^2_{Trad} = 2\sum_{q=1}^{Q} N\left[O_q Ln\frac{O_q}{E_q} + (1-O_q)Ln\frac{1-O_q}{1-E_q} \right] \qquad [4.19]$$

Si se cumple la Hipótesis nula (i.e., el modelo se ajusta a los datos) el estadístico anterior se distribuye según χ^2 con Q grados de libertad. Este estadístico está implementado en programas como BILOG o PARSCALE, pero su uso es desaconsejable si el test es corto (p.ej., menos de 20 ítems) ya que los valores pueden sobreestimarse si la agrupación de los evaluados por su nivel de rasgo no es precisa (ver por ejemplo, Stone y Zhang, 2003). En ese caso pueden aparecer discrepancias entre O_q y E_q que no se deben al desajuste al modelo.

Para resolver ese problema, Orlando y Thissen (2000) propusieron un estadístico para contrastar si la probabilidad de acertar observada como función de la puntuación observada X (O_X) difiere estadísticamente de la probabilidad de acertar según el modelo (E_X):

$$\chi^2_{Orlando} = I\sum_{X=1}^{J-1}\left[\frac{(O_X - E_X)^2}{E_X} + \frac{((1-O_X) - (1-E_X))^2}{1-E_X} \right] \qquad [4.20]$$

De esta manera no se requiere agrupar a los evaluados por una variable no observable, θ. El valor E_X se calcula mediante el algoritmo iterativo de Lord-Wingersky (1984) y requiere el uso de un programa informático para su obtención. Si se cumple la Hipótesis nula (el modelo se ajusta a los datos) el estadístico anterior se distribuye según χ^2 con $J - 1 - t$ grados de libertad, donde t es el número de parámetros estimados para el ítem.

En la actualidad, no es fácil decidir qué índices de bondad de ajuste son los mejores. Un problema de los contrastes estadísticos es que con muestras grandes las discrepancias pueden ser estadísticamente significativas, pero ser irrelevantes desde el punto de vista práctico. Lo contrario también puede ocurrir. Grandes discrepancias pueden no resultar estadísticamente significativas si la muestra es demasiado pequeña. Nuestra recomendación es completar la información de estos estadísticos con una inspección visual del ajuste de la CCI, tal como se hace en el siguiente ejemplo.

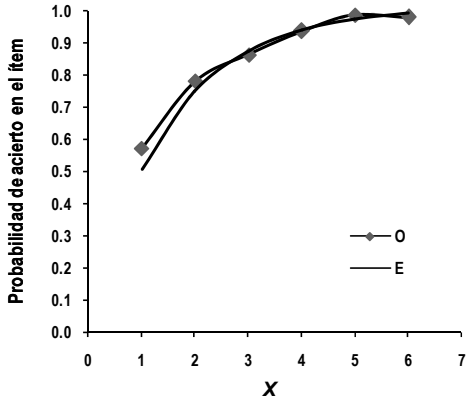
Ejemplo 4.6. Ajuste para los ítems del test de cálculo numérico

Para cada ítem de cálculo numérico se obtuvo el estadístico χ^2 de Orlando y Thissen. Este indicador se puede obtener con el programa GOODFIT de libre distribución (Orlando y Thissen, 2000). Los resultados se muestran en la Tabla 4.6 y la información gráfica para el ítem 1 en la Figura 4.12. Puede comprobarse que el ítem 1 muestra el peor funcionamiento. Sin embargo, la inspección visual permite comprobar que la diferencia entre la curva predicha por el modelo y la curva observada, aunque estadísticamente significativa, es irrelevante desde el punto de vista práctico.

Tabla 4.6. Índices de ajuste basados en la comparación de las probabilidades de acertar (observada y esperada) como función del test

Ítems	$\chi^2_{Orlando}$	*gl*	*P*
1	11,5	3	0,009
2	11,3	3	0,010
3	3,96	3	0,266
4	3,55	3	0,314
5	2,84	3	0,417
6	1,89	3	0,596
7	3,64	3	0,303

Figura 4.12. Probabilidades observada y esperada de acertar el ítem 1 como función de la puntuación *X*



La precisión de las puntuaciones en TRI

Función de información y error típico de estimación de θ

En la TCT un concepto fundamental es el error típico de medida (ver ecuación [3.37]) que nos permite conocer en qué grado la puntuación empírica en un test, *X,* es una buena aproximación a la puntuación verdadera, *V*. En concreto, a partir de S_E, se puede establecer el intervalo de confianza en torno al cual se encuentra la puntuación verdadera de una persona. En TRI, un concepto análogo al error típico de medida es el *error típico de estimación de* θ. Si aplicáramos un test con un suficiente número de ítems a personas con igual θ, la estimación ML de θ ($\hat{\theta}$) se distribuiría normalmente con media igual al parámetro verdadero (θ) y desviación típica $S_e(\theta)$, que es el error típico de estimación de θ:

$$Se(\theta) \equiv \sigma(\hat{\theta} \mid \theta) = \frac{1}{\sqrt{I(\theta)}} \qquad [4.21]$$

donde $I(\theta)$ se denomina *función de información del test.* Como se muestra en la ecuación, cuanto mayor sea la información, menor será el error típico de estimación. La función de información en TRI es un concepto análogo al de coeficiente de fiabilidad en TCT. Ambas son medidas de precisión a partir de las cuales se deriva un error típico (de medida en TCT y de estimación de θ en TRI). Las diferencias principales son que:

1. Mientras que el coeficiente de fiabilidad es un valor escalar, la función de información es una función; es decir, en TRI el valor de precisión varía para cada valor de θ.
2. Mientras que el coeficiente de fiabilidad puede tomar valores entre 0 y 1, la función de información puede tomar cualquier valor igual o superior a 0.

La función de información del test, $I(\theta)$, se obtiene como la suma de las funciones de informaciones de los ítems:

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta) \qquad [4.22]$$

La función de información de cada ítem para los modelos de uno, dos y tres parámetros se muestra en la Tabla 4.7, donde $P^*_j(\theta)$ se define en la ecuación [4.13] y $Q^*_j(\theta) = 1 - P^*_j(\theta)$.

Tabla 4.7. Ecuaciones para calcular la función de información de un ítem en los modelos logísticos

Modelo	*Ecuación para calcular la función de información*	
ML1P	$I_j(\theta) = D^2 a^2 P_j(\theta) Q_j(\theta)$	[4.23]
ML2P	$I_j(\theta) = D^2 a_j^2 P_j(\theta) Q_j(\theta)$	[4.24]
ML3P	$I_j(\theta) = D^2 a_j^2 P_j^*(\theta) Q_j^*(\theta)(1 - c_j)\left(P_j^*(\theta) / P_j(\theta)\right)$	[4.25]

Por tanto, el valor de la función de información del test dependerá de varios factores:

1. *Número de ítems aplicado* (como ocurría en la TCT): En general, al aumentar la longitud del test aumenta la información (ver ecuación [4.22]).
2. *De los parámetros a y c de los ítems aplicados:* a mayores parámetros de discriminación y menores parámetros de adivinación, mayor será $I(\theta)$ (ver ecuaciones [4.23] a [4.25]).
3. *De la proximidad entre θ y b_j:* cuanto menor sea la distancia entre los parámetros de dificultad de los ítems aplicados y el nivel de rasgo de la persona, mayor será $I(\theta)$. Los productos $P_j(\theta)Q_j(\theta)$, para el ML1P y el ML2P, y el producto $P^*_j(\theta)Q^*_j(\theta)$, para el

ML3P, alcanzan su máximo valor cuando $\theta = b_j$. Estos productos aparecen en las ecuaciones [4.23] a [4.25]

4. Del grado en que $P_j(\theta)$ se aleja de c_j: cuanto más próxima se encuentre la probabilidad a la esperada por efecto de la adivinación, menor será $I(\theta)$ (ver ecuación [4.25]; el cociente $P^*_j(\theta)/P_j(\theta)$ alcanza su valor máximo para niveles altos de θ, cuando $P^*_j(\theta)/P_j(\theta) \cong 1$).

Ejemplo 4.7. Función de información del test de cálculo numérico

En la Tabla 4.8 se muestra la información proporcionada, para distintos niveles θ, por los ítems de cálculo numérico y por el test completo.

Tabla 4.8. Función de información de los ítems y del test

Ítem	*a*	*b*	*c*	–3	–2	–1	0	1	2	3
1	0,68	–1,92	0,33	0,072	0,164	0,154	0,076	0,028	0,009	0,003
2	0,97	–0,97	0,10	0,021	0,207	0,555	0,336	0,088	0,018	0,003
3	0,51	0,20	0,11	0,013	0,043	0,099	0,147	0,141	0,094	0,049
4	0,59	–0,88	0,24	0,024	0,082	0,150	0,144	0,084	0,038	0,015
5	0,52	1,53	0,07	0,003	0,011	0,038	0,093	0,155	0,167	0,120
6	1,14	0,71	0,06	0,000	0,001	0,045	0,460	0,791	0,246	0,041
7	1,42	0,26	0,36	0,000	0,000	0,019	0,506	0,433	0,054	0,005
Test				0,133	0,509	1,059	1,763	1,721	0,626	0,237

Por ejemplo, la función de información del ítem 2 para $\theta = 0$. Se calcula como:

$$P_2^*(\theta) = \frac{1}{1+e^{-Da_2(\theta-b_2)}} = \frac{1}{1+e^{-1,702(0,97)(0-(-0,97))}} = 0,832$$

$$P_2(\theta) = c_2 + (1-c_2)\frac{1}{1+e^{-Da_2(\theta-b_2)}} = 0,10 + \frac{0,90}{1+e^{-1,702(0,97)(0-(-0,97))}} = 0,849$$

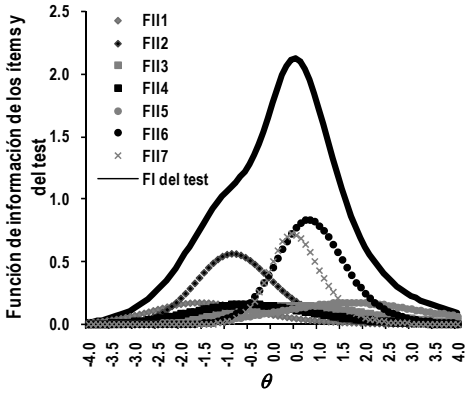
$$I_2(\theta) = D^2 a_2^2 P_2^*(\theta) Q_2^*(\theta)(1-c_2)\left(\frac{P_2^*(\theta)}{P_2(\theta)}\right) =$$

$$= 1,702^2 0,97^2 0,832(1-0,832)(1-0,10)\left(\frac{0,832}{0,849}\right) = 0,336$$

Puede observarse que el test proporciona la mayor información para los niveles de rasgo entre 0 y 1.

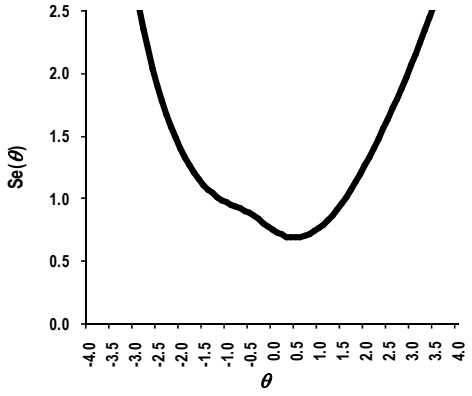
Normalmente, la función de información se representa de forma gráfica. En la Figura 4.13 se observa que los que más contribuyen a la precisión son los ítems 2, 6 y 7 (los más discriminativos). Los ítems 1, 3, 4 y 5 proporcionan muy poca información (i.e., sus funciones de información son bastante planas). Para aumentar la precisión en niveles de θ bajos (p.ej., menores que cero) deberíamos añadir ítems similares en dificultad al ítem 2.

Figura 4.13. Función de información de los ítems y del test



El error típico de estimación se representa en la Figura 4.14. Observe que el error típico y la información están inversamente relacionados. Cuando la información es mayor, el error típico es menor, y viceversa. Puede concluirse que, en general, la precisión del test no es adecuada, especialmente a la hora de discriminar entre niveles de rasgo bajos o entre niveles de rasgo altos[7].

Figura 4.14. Error típico de estimación de θ



[7] Debe observarse que la función de información depende del modelo aplicado. Por ejemplo, en niveles de θ bajos, la aplicación del ML1P dará valores mayores que el ML3P en la función de información, $I(\theta)$. Sin embargo, los distintos modelos (ML1P, ML2P, ML3P) no deben compararse en este sentido. Si el modelo de un parámetro no se ajustara a los datos, las fórmulas que habríamos proporcionado para obtener el error típico estimación dejarían de ser válidas.

El hecho de que la función de información sea la suma de las funciones de información de los ítems nos permite elegir los ítems más adecuados en cada momento en función de las demandas de la aplicación. Por ejemplo, si en un proceso de selección de personal sólo vamos a elegir unos pocos evaluados muy competentes, a partir de un banco de ítems calibrado podríamos elegir aquellos que proporcionan más información para niveles altos de θ. Esto nos permitiría aplicar un número reducido de ítems sin perder precisión al estimar θ.

En general, un ítem j es máximamente preciso para niveles de rasgo $\theta = b_j$ (en el caso del ML1P y del ML2P) o, de forma más general, cuando $\theta = \theta_{max}$, siendo θ_{max} (Hambleton, Swaminathan y Rogers, 1991; p. 92):

$$\theta_{\max} = b_j + \frac{\ln(0{,}5 + 0{,}5\sqrt{1 + 8c_j})}{Da_j} \qquad [4.26]$$

que es valor de rasgo para el cual el ítem proporcionará la información máxima; esta información máxima puede calcularse de la siguiente forma (Hambleton y Swaminathan, 1985):

$$I(\theta_{\max}) = 0{,}25D^2a_j^2 \frac{\left[1 - 20c_j - 8c_j^2 + (1 + 8c_j)^{3/2}\right]}{2(1 - c_j)^2} \qquad [4.27]$$

expresión que, en el ML2P, se reduce a $I(\theta_{max}) = 0{,}25D^2a_j^2$.

Intervalos de confianza para la estimación de θ

A partir del error típico de estimación se puede obtener el intervalo confidencial en el que, con probabilidad predeterminada, se ha de encontrar el nivel de rasgo de la persona. En concreto, si al nivel θ estimado de una persona le sumamos y restamos $(1{,}96)S_e(\theta)$, obtenemos los extremos del intervalo en el que, con probabilidad 0,95, se encontrará su verdadero nivel de rasgo. Por ejemplo, si la θ estimada es 0,8 y su error típico de estimación es 0,22, entonces, el nivel de rasgo de dicha persona se encuentra entre 0,37 (pues 0,8 – (1,96)0,22 = 0,37) y 1,23 (pues 0,8 + (1,96)0,22 = 1,23), con probabilidad 0,95.

Función de información y fiabilidad

A partir de la función de información del test se puede obtener un *coeficiente de fiabilidad marginal* para las estimaciones del nivel de rasgo:

$$r_{\hat{\theta}\hat{\theta}}^{TRI} = \frac{\sigma_{\theta}^2}{\sigma_{\hat{\theta}}^2} = \frac{\sigma_{\theta}^2}{\sigma_{\hat{\theta}}^2 + \sum_{q}^{Q} Se(\theta_q)^2 \, g(\theta_q)} \qquad [4.28]$$

La expresión $g(\theta)$ indica la distribución del rasgo. $r_{\theta\theta}^{TRI}$ es el cociente entre la varianza del verdadero nivel de rasgo, σ^2_{θ}, y la varianza del rasgo estimado, $\sigma_{\hat{\theta}}^2$; indica qué proporción de la varianza de las estimaciones es varianza verdadera. Mediante el uso de esta fórmula puede anticiparse el coeficiente de fiabilidad que se obtendría en una muestra en la que el rasgo tenga distribución $g(\theta)$ y varianza σ^2_{θ}. Otra forma de expresar el coeficiente de fiabilidad marginal es como:

$$r_{\hat{\theta}\hat{\theta}}^{TRI} = \frac{\sigma_{\hat{\theta}}^2 - \sum_{q}^{Q} Se(\theta_q)^2 \, g(\theta_q)}{\sigma_{\hat{\theta}}^2} \qquad [4.29]$$

Si $\hat{\theta}$ está estandarizada la ecuación se simplifica a:

$$r_{\hat{\theta}\hat{\theta}}^{TRI} = 1 - \sum_{q}^{Q} Se(\theta_q)^2 \, g(\theta_q)$$

En ocasiones, también se calcula cuál sería el coeficiente de fiabilidad si todos los evaluados de una muestra fueran medidos con la precisión que se obtiene en un nivel de θ dado ($\theta = \theta_q$). En ese caso, se aplica la siguiente fórmula:

$$r_{\hat{\theta}\hat{\theta}}^{TRI}(\theta_q) = \frac{\sigma_{\hat{\theta}}^2 - Se(\theta_q)^2}{\sigma_{\hat{\theta}}^2} \qquad [4.30]$$

Si $\hat{\theta}$ está estandarizada la ecuación se simplifica a:

$$r_{\hat{\theta}\hat{\theta}}^{TRI}(\theta_q) = 1 - Se(\theta_q)^2$$

La Curva Característica del Test (CCT)

La representación de la relación entre θ y el rendimiento esperado en el test se denomina como *Curva Característica del Test*. Para un valor θ concreto, el valor esperado en el test se obtiene como la suma de las correspondientes probabilidades de acierto de los ítems para dicho nivel de rasgo, que pueden obtenerse en las correspondientes curvas características:

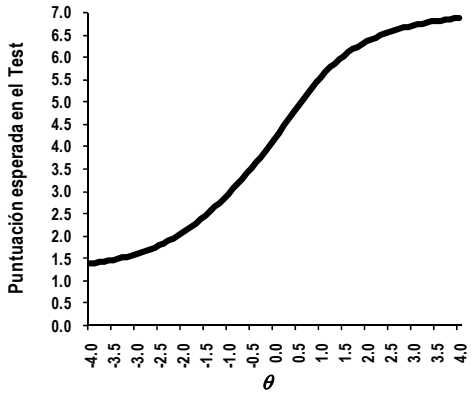
$$X(\theta) \equiv \varepsilon(X \mid \theta) = \sum_{j=1}^{J} P_j(\theta) \qquad [4.31]$$

La CCT permite la transformación de la escala de θ a la escala de puntuaciones directas. Además la CCT juega un papel importante en algunos de los procedimientos de equiparación de parámetros (tal como se verá en el capítulo 11).

Ejemplo 4.8. Curva Característica del Test

La CCT del test de 7 ítems de cálculo numérico se representa en la Figura 4.15. Puede observarse que la relación entre θ (el nivel de rasgo) y X (la puntuación esperada en el test) no es lineal. Para alguien con un nivel de θ de 4 se espera un número esperado de aciertos próximo al número de ítems (en nuestro caso, 7).

Figura 4.15. CCT del test de cálculo numérico



Aplicaciones

El desarrollo de la TRI ha supuesto un cierto avance, tanto en ciertos contextos aplicados de evaluación psicológica y educativa, como en contextos de investigación muy diversos. Tal como vamos viendo, aplicar la TRI a las puntuaciones que se obtienen en los ítems de un test tiene ciertas ventajas, siendo una de las principales la estimación del error que se comente concretamente con cada persona. Los desarrollos de la TRI permiten aplicaciones más eficientes, ya que facilitan el ensamblado de un test (la selección de los ítems que lo formarán) para optimizar la precisión de las estimaciones de rasgo. También permite obtener indicadores psicométricos para los ítems, complementarios (y relacionados) a los

de la TCT. El estudio de los parámetros de los ítems que se estiman en diferentes grupos va a facilitar el análisis de posibles problemas no deseados, como sería que la prueba perjudicara a uno de dicho grupos sin motivo justificado.

Este marco teórico no resulta, sin embargo, la panacea universal para analizar las mediciones que se realizan con cualquier tipo de test en cualquier tipo de contexto de evaluación. Resultando muchas veces complementaria a la TCT, la TRI proporciona su mayor utilidad en los estudios de evaluación a gran escala, donde es preciso medir a muestras numerosas de personas y en diferentes ocasiones. Por una parte, en este tipo de estudios (p.ej., sobre evaluación educativa) se satisfacen los requisitos muestrales necesarios para su aplicación; por otra parte, en aplicaciones transculturales de tests se demandan ciertos estudios particulares (p.ej., equiparación de puntuaciones o estudio del funcionamiento diferencial de los ítems) para los que la TRI proporciona métodos y procedimientos más adecuados que la TCT.

Algunas de las principales aplicaciones de la TRI se describen con cierto detalle en otros capítulos de este libro: Equiparación de Parámetros (capítulo 11), Funcionamiento Diferencial (capítulo 13) y Tests Adaptativos Informatizados (capítulo 15).

Ventajas y desventajas de los modelos de TRI

Las aplicaciones de la TRI no serían posibles sin ciertas propiedades teóricas de estos modelos, que sintetizamos en las siguientes:

1. *Invarianza de los parámetros de los evaluados.* En TRI se concibe de forma más operativa el nivel de rasgo del evaluado. Desaparece el concepto de puntuación verdadera, que en la TCT se ligaba íntimamente al test utilizado (por ejemplo, la puntuación verdadera cambiaba si el test tenía más ítems o ítems con distinta dificultad). La TRI se centra en las propiedades psicométricas de los ítems y, a partir de ellas, deriva las propiedades psicométricas del test. En TRI se establece que el nivel de rasgo es un parámetro θ del evaluado que puede ser estimado una vez se conocen los parámetros de los ítems del test que se está aplicando.
   No debe confundirse la invarianza de parámetros con la invarianza de las estimaciones de los parámetros; es decir, el parámetro de la persona que indica su nivel de rasgo será el mismo se utilice un test corto o un test largo, pero la estimación de ese parámetro dependerá del test utilizado (se realizará con más precisión en el test largo).
2. *Invarianza de parámetros de los ítems.* Si se cumplen los supuestos del modelo, los parámetros estimados de los ítems no dependen, salvo transformación lineal (ver Apéndice), de la muestra donde se obtienen. Esto permite que podamos estudiar las propiedades psicométricas del mismo ítem en distintos grupos, incluso si difieren en media o variabilidad en el nivel de rasgo.
3. *No se asume la homocedasticidad del error.* En TRI la precisión del test está condicionada al nivel de rasgo y a los ítems aplicados. De esta manera, se esquiva una de las principales críticas que se ha realizado al concepto de fiabilidad en TCT.
4. *Se dispone de indicadores de bondad de ajuste,* que hacen el modelo falsable y permiten así la comparación de distintos modelos alternativos para las respuestas.

Sin embargo, no todo son ventajas en el actual desarrollo que tienen los principales modelos de la TRI:

1. Se requiere un gran número de evaluados para obtener las estimaciones de los parámetros, especialmente en los modelos más complejos (p.ej., los que incluyen parámetros de adivinación o algunos modelos para ítems politómicos) (Thissen y Wainer, 1982). Como el problema de los requisitos muestrales no es independiente del método de estimación empleado, se tratará con más detalle en el capítulo 11.
2. Los supuestos son muy restrictivos. Los modelos de Rasch son los más afectados por esta crítica, ya que asumen ausencia de adivinación en las respuestas (algo poco razonable cuando trabajamos con ítems de opción múltiple) e igual parámetro de discriminación de los ítems (algo que, por lo general, no se cumple). Los supuestos de unidimensionalidad o independencia local pueden ser poco realistas en algunos casos. Afortunadamente, se están desarrollando modelos psicométricos de TRI que no requieren el cumplimiento de estos supuestos (p.ej., modelos multidimensionales o modelos para testlets).
3. Los procedimientos para comprobar el ajuste no son totalmente satisfactorios, fundamentalmente porque se desconoce la distribución de los índices de ajuste. Por ejemplo, algunos indicadores dependen de la longitud del test o de la calidad de las estimaciones del nivel de rasgo.
4. La concepción sobre las fuentes de error que afectan a las puntuaciones de las personas en los tests es limitada, sobre todo si la comparamos con la propuesta que se hace desde la Teoría de la Generalizabilidad (TG). La mayor parte de los modelos de TRI ignoran aquellas fuentes de error de medida que no están relacionadas con el contenido específico de los ítems. La TG permite el estudio del efecto de diferentes fuentes de error, tal como veremos en el capítulo 9. En TRI no se reconocen distintas fuentes de error (Brennan, 2004), aunque algunos tímidos intentos empiezan a esbozarse (Bock, Brennan y Muraki, 2002).
5. Como también ocurre en TCT, la TRI se centra en el problema de la precisión, con lo que ha desviado de algún modo la atención de los psicómetras hacia problemas técnicos (p.ej., la estimación de parámetros o la evaluación del ajuste), ignorando en parte el tema de la validez (Muñiz, 1996). La TRI (al menos los modelos descritos hasta el momento) es fundamentalmente una teoría descriptiva (no psicológica) sobre el modelo de respuesta a los ítems.

Apéndice

La escala métrica de θ

Para comparar dos objetos en un atributo necesitamos que las medidas se hayan tomado en la misma escala. El concepto de escala métrica hace referencia a las unidades de medida y al origen (i.e., el punto cero) de una escala. Por ejemplo, la temperatura puede expresarse en una escala de grados Celsius o en una escala de grados Fahrenheit; ambas son legitimas y equivalentes mediante la correspondiente transformación lineal [$Temp(F^o) = 1{,}8Temp(C^o) + 32$**]**.

Para las puntuaciones en un test se suele asumir un nivel de medida de intervalo, como para la temperatura. Esto quiere decir que no hay un cero absoluto en la escala que indique un nivel cero de atributo y, por lo tanto, no sólo son arbitrarias las unidades de medida (i.e.: hablar en F^o o en C^o) sino también el origen de la escala. En este nivel de medida, como no existe un cero absoluto, la afirmación de que un objeto tiene el doble de temperatura que otro no tiene sentido; como tampoco lo tiene, por ejemplo, afirmar que una persona es el doble de inteligente que otra. En otras palabras, el nivel de medida determina qué transformaciones de la escala son posibles y qué afirmaciones acerca del atributo tienen sentido y cuáles no.

¿Qué valores puede tomar θ? ¿Cuál es el origen o punto cero de la escala? ¿Y las unidades de medida? Al ser θ una variable con nivel de medida de intervalo, el origen es arbitrario. En la práctica, suele trabajarse con la escala θ en puntuaciones típicas ($\mu_\theta = 0$; $\sigma^2_\theta = 1$). Esto quiere decir que la escala de θ es tal que, en la muestra, la media es cero y la varianza 1; los valores θ variarán generalmente entre −3,0 y 3,0. Una vez definida la escala para θ, automáticamente los parámetros a y b se sitúan en una métrica consistente con ella: por ejemplo, los valores del parámetro b variarán, generalmente, entre −3,0 y 3,0; el valor de a entre 0,3 y 2,5. A lo largo del capítulo se ha asumido esta escala métrica para θ.

Sin embargo, el nivel θ del evaluado puede definirse en cualquier escala métrica consistente con su nivel de medida, en este caso de intervalo. Esto quiere decir que si efectuamos una transformación lineal de θ ($g > 0$):

$$\theta^* = g\theta + h \qquad [4.32]$$

las probabilidades de acierto no cambian si, a la vez, transformamos también los parámetros a y b de los ítems para que se hallen en una métrica consistente con la de θ^*:

$$a^* = \frac{a}{g} \qquad [4.33]$$

$$b_j^* = gb_j + h \qquad [4.34]$$

En efecto:

$$P_j(\theta^*) = \frac{1}{1+e^{-Da^*(\theta^*-b_j^*)}} = \frac{1}{1+e^{-D\frac{a}{g}(g\theta+h-(gb_j+h))}} = \frac{1}{1+e^{-Da(\theta-b_j)}} = P_j(\theta)$$

Por tanto, el modelo con parámetros θ^*, a^* y b^*_j es equivalente al modelo con parámetros θ, a y b_j. Es decir, lo mismo que en la temperatura, podemos expresar los parámetros en distinta escala. El rango de valores que pueden tomar los parámetros a, b y θ dependerá de la escala métrica que utilicemos. Para resolver esta indeterminación debemos explicitar si θ está en una escala de puntuaciones típicas o en otra escala.

Observe que ahora podemos explicar por qué las ecuaciones del ML1P [4.1] y del modelo de Rasch [4.2] son equivalentes. En efecto, si definimos:

$$b^*_j = Dab_j \qquad [$$
$$\theta^* = Da\theta$$

Entonces el modelo de Rasch se transforma en el ML1P:

$$\frac{1}{1+e^{-(\theta^*-b^*_j)}} = \frac{1}{1+e^{-(Da\theta-Dab_j)}} = \frac{1}{1+e^{-Da(\theta-b_j)}}$$

Si el parámetro θ del ML1P se expresa en puntuaciones típicas ($\sigma_\theta = 1$), entonces la desviación típica del parámetro θ^* en el modelo de Rasch será:

$$\sigma_{\theta^*} = Da$$

Otro ejemplo de esta necesidad de explicitar la escala métrica se relaciona con el parámetro de discriminación de los ítems: el parámetro a de un modelo será distinto si utilizamos $D = 1{,}702$ o $D = 1$. La elección de uno u otro define la escala métrica de la discriminación: parámetro a en métrica normal o en métrica logística.

La escala *logit*

Si p es una probabilidad, la función *logit* de p es ln[p/(1 p)]

$$\text{logit}(p) \equiv \ln\left[\frac{p}{1-p}\right]$$

En el modelo de Rasch suele utilizarse la escala *logit* (*log-odds*-unit) para informar de los parámetros. Al utilizar la ecuación [4.2], el *logit* de la probabilidad de acertar un ítem es:

$$\ln\left[\frac{P_j(\theta)}{Q_j(\theta)}\right] = \ln\left(\frac{\dfrac{1}{1+e^{-(\theta-b_j)}}}{1-\dfrac{1}{1+e^{-(\theta-b_j)}}}\right) = \ln\left(e^{(\theta-b_j)}\right) = \theta - b_j$$

Es decir, depende sólo del nivel de rasgo y de la dificultad del ítem. Utilizando esta escala es más fácil interpretar las diferencias de rendimiento entre dos personas en el mismo ítem o de la misma

persona en dos ítems. Por ejemplo, una diferencia en la escala *logit* de las probabilidad de acertar un ítem que tienen dos personas , será:

$$\ln\left[\frac{P_j(\theta_2)}{Q_j(\theta_2)}\right] - \ln\left[\frac{P_j(\theta_1)}{Q_j(\theta_1)}\right] = (\theta_2 - b_j) - (\theta_1 - b_j) = \theta_2 - \theta_1$$

Lo que muestra dicha diferencia no depende del ítem, sino únicamente de los dos niveles de rasgo. Una diferencia en la escala *logit,* para una persona, entre las probabilidades de acertar dos ítems distintos, será:

$$\ln\left[\frac{P_2(\theta)}{Q_2(\theta)}\right] - \ln\left[\frac{P_1(\theta)}{Q_1(\theta)}\right] = (\theta - b_2) - (\theta - b_1) = b_2 - b_1$$

En el modelo de Rasch, por tanto, las diferencias en la escala *logit* se corresponden directamente con las diferencias en θ (o en b). Por tanto, al informar del nivel de rasgo en la escala *logit* se informa de θ y al informar de la dificultad en la escala *logit* se informa de b. El punto cero de la escala *logit* es arbitrario. Normalmente se establece como punto cero la media de los parámetros b de los ítems o la media del nivel de rasgo. En el primer caso, los valores θ iguales a 0 indican que se tiene una probabilidad de acertar ítems de dificultad media de 0,5. En el segundo caso, los valores b iguales a 0 se corresponden con ítems que los evaluados de nivel medio aciertan con probabilidad 0,5.