Revista de Psicología Clínica con Niños y Adolescentes
Vol. 3. Nº. 2 - Julio 2016 - pp 13-21

www.revistapcna.com - ISSN 2340-8340

Revista de Psicología Clínica con Niños y Adolescentes

Una introducción didáctica a la Teoría de Respuesta al Ítem para comprender la construcción de escalas

María D. Hidalgo-Montesinos[1] & Brian F. French[2]
[1]*Universidad de Murcia, España*
[2]*Washington State University, Estados Unidos*

Resumen

En este trabajo se ofrece una introducción a la Teoría de Respuesta al Ítem (TRI), que proporcionará al lector una visión general de las ideas fundamentales que subyacen a estos modelos y cómo el análisis de ítems usando estos modelos ayuda en el proceso de construcción del test. La TRI representa una alternativa a la Teoría Clásica de Tests (TCT). La TCT tiene una larga historia de uso en la medición psicológica y educativa, particularmente en el desarrollo de escalas de personalidad. Sin embargo, en muchos casos TRI y TCT se combinan para el desarrollo de escalas. Se explican los modelos y procedimientos de la TRI que nos permiten examinar el comportamiento de los ítems del test. Estos conceptos se aplican a una escala de depresión-ansiedad utilizada en estudiantes de Educación Secundaria para identificar riesgos y necesidades y ayudar en la intervención psicológica-educativa. Se analizan los ítems de esta escala según el Modelo de Respuesta Graduada, que resulta apropiado para ítems de respuesta ordinal. Este trabajo se centra en cómo se puede utilizar la información que proporcionan estos modelos para asegurar que los ítems se ajustan al propósito para el que fueron diseñados, es decir, al rasgo que pretenden medir (depresión-ansiedad). Tanto los resultados estadísticos como la información gráfica obtenida se muestran como apoyo para la comprensión de los conceptos básicos en TRI y de la profundidad de la información que tales análisis proporcionan. Por último, se ofrece información sobre software y recursos disponibles para el análisis usando TRI.
Palabras clave: teoría de respuesta al ítem, construcción de tests, modelo de respuesta graduada.

Abstract

A didactic introduction to Item Response Theory for understanding the construction of scales. This article offers a gentle introduction to Item Response Theory (IRT). This introduction will provide the reader with a broad overview of key ideas underlying IRT and how IRT analyses can be used to aid one's work. IRT represents an alternative to classical test theory (CTT). CTT has a long history of use in the area of educational and psychological measurement and psychometrics, particularly for the development of personality scales. However, in many instances IRT is used in combination with CTT for scale development. Thus, IRT is introduced in the context of the scale development process. Specifically, IRT models and methods are explained through the examination of the behavior of items that comprise a scale. The IRT concepts are applied to a depression-anxiety scale used with students in secondary education to identify risk and needs to aid intervention. Items are evaluated with an empirical item analysis to demonstrate the basic IRT model, the Graded Response Model, for ordinal level item responses. The demonstration focuses on how such information can be utilized to ensure items meet the purpose of the scale in relation to the trait (i.e., depression-anxiety) measured. Both statistical and graphical information are demonstrated to aid in the understanding of IRT concepts and the depth of information such analyses provide. Finally, advice about software and resources available for IRT analysis is offered. This introduction should increase the reader's knowledge of IRT. Moreover, the reader will become a more critical and informed consumer of test development Information.
Keywords: item response theory, test development, graded response model.

En evaluación psicológica y de la salud los tests, cuestionarios o escalas utilizados para la toma de decisiones acerca de un individuo o grupo deben ser construidos y evaluados de manera apropiada y considerando el uso previsto de los mismos. Los tests pueden servir para varios propósitos, dado que las diferentes evidencias de validez que se aportan apoyan las decisiones a tomar. Tanto el investigador o desarrollador de un instrumento como el profesional que selecciona un test para su uso en evaluación debe estar familiarizado con las directrices y normas establecidas por varios organismos y comisiones, como las Directrices para el Uso de los Tests de la Comisión

Correspondencia:
María D. Hidalgo-Montesinos.
Departamento de Psicología Básica y Metodología. Facultad de Psicología.
Universidad de Murcia. Campus de Espinardo, Apdo. 4021, C. P. 30100, Murcia, España.
E.mail: mdhidalg@um.es

Internacional de Tests (ITC, 2001) o los *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Estos documentos proporcionan una excelente guía respecto a todas las fases en el proceso de evaluación. En este trabajo, nos centraremos en uno de los pasos del proceso de desarrollo de un instrumento de evaluación, en concreto el referido al análisis de la utilidad de los ítems para el propósito de la evaluación y en qué medida estos ítems permiten cumplir con los objetivos de evaluación definidos en el propio test. Para tal fin introduciremos al lector en los modelos de medida de la Teoría de Respuesta al Ítem (TRI) y expondremos un ejemplo a través del análisis de una escala de *depresión-ansiedad* para su uso con estudiantes en Educación Secundaria.

Teoría de Respuesta al Ítem: supuestos y modelos

La Teoría de Respuesta al Ítem representa una alternativa a la Teoría Clásica de Tests (TCT). Sin duda, la TCT es más popular en su uso, de este modo, la mayoría de los instrumentos que se han desarrollado lo han sido a partir de la misma. Sin embargo, no podemos obviar que hemos sufrido un cambio importante y se ha pasado a la utilización de modelos matemáticos de medida que imponen severas restricciones a los datos para justificar que el instrumento construido y los ítems que forman parte del mismo miden de manera apropiada el constructo o variable de interés. Sin embargo, observamos que en muchos casos se combina el uso de ambos métodos, TRI y TCT, en el desarrollo de tests (Muñiz, 2010). Los modelos de respuesta al ítem, al igual que los modelos de regresión logística, se pueden considerar modelos de regresión no lineal (Hambleton, Swaminathan, & Rogers, 1991; Neter, Kutner, Nachtsheim, & Wasserman, 1996). El análisis usando TRI, a través de modelos matemáticos, nos proporciona una visión de la relación entre el nivel en el rasgo de un individuo (por ejemplo, nivel de depresión) y las características de los ítems. La TRI depende de algunos supuestos clave. Dos supuestos importantes de estos modelos son unidimensionalidad e independencia local de los ítems. El primero supone que los ítems miden esencialmente uno y sólo un rasgo latente, mientras que el segundo, independencia local, asume que las respuestas de un individuo para cualquier pareja de ítems en el test no están relacionadas cuando consideramos un mismo nivel en el rasgo, es decir, cuando la habilidad se mantiene constante. Además, como característica importante a considerar, la TRI proporciona estimaciones invariantes de las propiedades psicométricas de los ítems, así como de las características de los sujetos, es decir, que los parámetros que caracterizan al ítem y al test son menos dependientes de la muestra particular de sujetos utilizada y que los parámetros que caracterizan al sujeto no dependen de la muestra particular de ítems utilizada. Este supuesto es una ventaja de la TRI que la hace especialmente recomendable debido a que los parámetros de los ítems se supone que son invariantes en la población de sujetos. Embretson y Reise (2009), Hambleton et al. (1991), López-Pina (1995), Muñiz (1997) y De Ayala (2009) sirven como excelentes fuentes para un tratamiento más en profundidad de los supuestos, también se puede consultar el libro de Abad, Olea, Ponsoda y García (2011).

La TRI proporciona una amplia gama de modelos que permite trabajar con tests tanto unidimensionales como multidimensionales y con distintos formatos de respuesta (dicotómico, politómico, continuo,...). En Mellenbergh (1994) y Hambleton y Van der Linden (1997) encontramos una exposición detallada de estos modelos. Estos modelos se diferencian en función del número de parámetros que contienen dependiendo de los supuestos que subyacen a los datos. Comenzaremos con un modelo que permite comprender fácilmente la idea de la TRI. El modelo logístico de 3 parámetros (3PL) (Hambleton et al., 1991; Lord & Novick, 1968) es un modelo muy popular (común) que se utiliza con ítems de respuesta dicotómica. Estos ítems son los típicos en un examen de rendimiento académico donde cada ítem se formula con dos opciones de respuesta, donde hay una respuesta que es la correcta y otra que es incorrecta. Un modelo de TRI predice la probabilidad de respuesta a un ítem basándose en diferentes parámetros de los ítems. En el modelo 3PL son tres los parámetros que definen las características de cada ítem (a) discriminación del ítem (es decir, parámetro *a*; un parámetro que mide la capacidad del ítem para diferenciar a los sujetos en función de su nivel en el rasgo latente), (b) la dificultad del ítem (es decir, parámetro *b*), y (c) pseudo-adivinación (es decir, parámetro *c*; indica la posibilidad de que un sujeto pueda acertar el ítem por azar). Además de los parámetros de los ítems, en cualquier modelo de TRI tenemos que considerar los parámetros referidos a los sujetos, es decir, los parámetros de habilidad o rasgo latente (θ). Otros modelos (por ejemplo, el de dos parámetros 2PL) se pueden utilizar, dependiendo de la naturaleza de los datos. Por ejemplo, si la adivinación no es posible o no se puede asumir, el 2PL (es decir, sin el parámetro *c*) puede ajustarse mejor a los datos que el modelo 3PL. El modelo 1PL, es un modelo bastante elegante y simple, es posible ajustarlo cuando se asume que no hay azar en las respuestas y los ítems presentan la misma capacidad discriminativa, es decir, igual discriminación. Por lo tanto, sólo el parámetro *b* se utiliza para predecir la probabilidad de una respuesta correcta. El modelo de 3PL viene dado por la siguiente expresión:

$$P(U_i = 1 \mid \theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}$$

donde

θ = *Habilidad o rasgo latente del evaluado*

c_i = *parámetro de pseudo - azar para el ítem i*

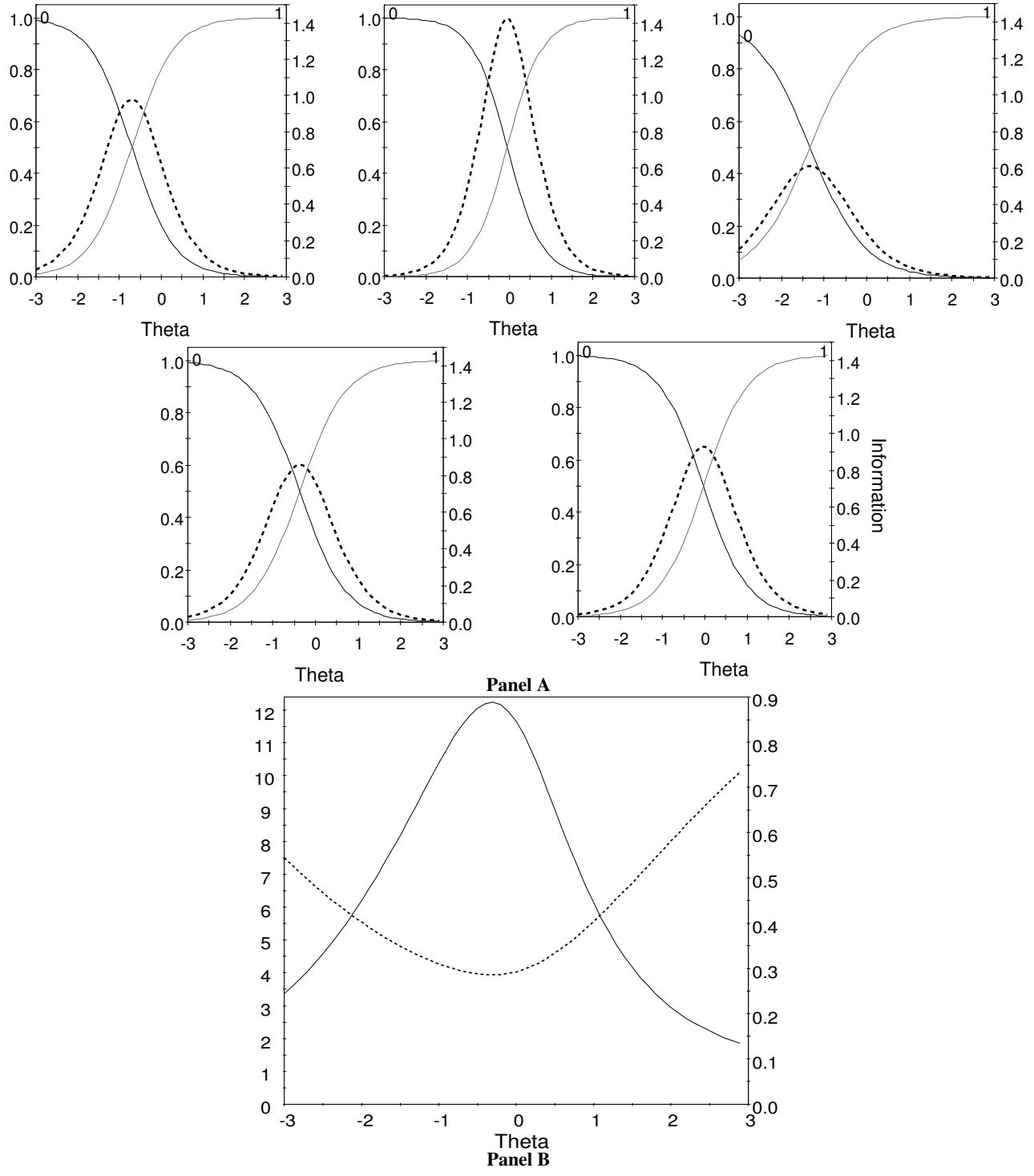
a_i = *parámetro de discriminación para el ítem i*

b_i = *parámetro de dificultad para el ítem i*

El modelado de la respuesta a un ítem proporciona lo que se denomina una función de respuesta al ítem (FRI) o curva característica del ítem (CCI). La Figura 1 contiene para cada uno de 5 ítems de respuesta dicotómica su CCI. Este gráfico ayuda a explicar la relación entre la habilidad latente que está siendo evaluada (conocimiento en matemáticas) y la probabilidad de un estudiante de responder correctamente al ítem. La dificultad del ítem (*b*) representa el nivel de dificultad del ítem y se define como el nivel de habilidad (a menudo etiquetado como θ, theta) en la que un individuo tiene una probabilidad del 50% de responder correctamente al ítem. La discriminación (*a*) representa la capacidad del ítem para discriminar entre individuos con diferentes niveles de habilidad, y es proporcional a la pendiente de la CCI en el valor de dificultad del ítem. Las CCIs del ejemplo, representan ítems con diferentes valores de discriminación, lo que indica que se puede diferenciar bastante bien a aquellos sujetos con niveles de habilidad cercanos a la dificultad del ítem tanto por encima de su valor como por debajo. Cuanto más pronunciada es la pendiente de la curva mayor es la discriminación del ítem. Por otro lado, cuando la curva para una respuesta de 1 (es decir, una respuesta correcta) se encuentra más desplaza a la derecha en la escala de habilidad, más difícil es el ítem, es decir, es necesario un nivel en el rasgo más alto para que el individuo acierte el ítem. Por ejemplo, el tercer ítem de la fila superior, más a la derecha, es el ítem más fácil, mientras que el ítem de la fila inferior de la derecha es el más

difícil. A través de una transformación de las CCIs podemos obtener funciones de información del ítem. En la Figura 1 (Panel A) las FFIs están representadas por las líneas de puntos. Tal y como se puede observar, el segundo ítem de la fila superior es el que proporciona la mayor información (es decir, el pico más alto), mientras que el tercer ítem de la fila superior proporciona menor cantidad de información (es decir, la curva con el pico más bajo). Estas funciones representan la cantidad de información que proporciona un ítem en cada nivel del continuo de habilidad. Sumando las FFIs de los ítems, obtenemos una función de información para el test de 5 ítems (parte inferior de la Figura 1, línea continua), que nos informa en qué nivel de habilidad la evaluación es más precisa. Esta función de información en TRI está relaciona con la fiabilidad de la medida. Cuanta más información tiene un ítem o un test, más precisa es la estimación de la habilidad para un sujeto. En TRI cuanto más precisa sea la estimación mayor será la fiabilidad. Una diferencia clave en TRI, en comparación con TCT, es que la información, y por lo tanto la precisión, puede variar a través de la distribución o continuo de habilidad. Además, el error típico (línea de puntos en la Figura 1, panel inferior) es la inversa de la función de información. El gráfico inferior de la Figura 1 (panel B) indica que este conjunto de ítems es más preciso cuando $\theta = -0.5$, y menos preciso en valores de habilidad por debajo de $\theta = -3$ o por encima de $\theta = 3$.

Figura 1. Panel A: Curvas Característica del Ítem para 5 ítems y sus respectivas funciones de información (línea de puntos) y Panel B: Función de Información del Test (línea continua) y Error Típico de Medida (línea discontinua).



Esta información, en combinación con las CCIs se puede utilizar para llevar a cabo un análisis empírico de los ítems y ayudar al desarrollo y depuración del instrumento de evaluación. Es decir, podemos utilizar esta información gráfica para señalar un lugar en la distribución de habilidad donde necesitamos más precisión y seleccionar ítems que permitan alcanzar ese objetivo en esa zona. Esto es muy útil cuando tenemos que ser precisos en un cierto nivel de habilidad para tomar decisiones que implicarán graves consecuencias sobre los individuos. Por ejemplo, si estamos seleccionando estudiantes para un programa de altas capacidades, necesitaremos una evaluación muy precisa en el nivel de habilidad de 2.5. Aquí es donde queremos que la curva de información alcance un valor más alto y, por tanto, el error típico sea el más bajo. Por lo tanto, nos gustaría buscar ítems que tengan FFIs con máximos cercanos a esta zona para asegurar que construimos un criterio preciso focalizando la evaluación en este nivel de habilidad. Esta situación también es la típica que se produce cuando tenemos que realizar un diagnóstico clínico usando un punto de corte en un test.

Una vez presentados los conceptos básicos de los modelos de TRI para ítems dicotómicos, vamos a pasar a modelos de TRI más complicados que subyacen a la mayoría de las escalas y cuestionarios de personalidad que pueden ser utilizados con adultos, adolescentes y niños en contextos de evaluación psicológica. Muchas de las medidas de personalidad utilizan ítems que se puntúan en una escala ordinal o de valoración (por ejemplo, *Totalmente de acuerdo* a *Totalmente en desacuerdo*), que solemos referirnos como ítems politómicos. Al igual que con los ítems puntuados de forma dicotómica, encontramos varios modelos de TRI que pueden ser seleccionados en función del tipo de datos, es decir, de los supuestos de esos datos. Los tres elementos claves para identificar el modelo de TRI más apropiado para tales tipo de datos son: (a) atractivo teórico, (b) tamaño apropiado de la muestra, y (c) ajuste del modelo (Penfield, 2014). En Penfield (2014) se presenta una excelente visión general de los modelos de TRI para ítems politómicos. Tales modelos incluyen el Modelo de Respuesta Graduada, el Modelo de Crédito Parcial, y el Modelo Nominal, por nombrar algunos. Sobre la base de las características de la escala de depresión-ansiedad que utilizaremos de ejemplo en este artículo, el modelo más apropiado sería el Modelo de Respuesta Graduada (MRG). El MRG es un modelo comúnmente usado, ya que se puede aplicar a escalas que utilizan varias opciones de respuesta. Este modelo implica que las puntuaciones en un ítem (por ejemplo, 0, 1, 2, 3) están ordenadas, y una puntuación o respuesta más alta, de un sujeto indica un nivel más alto en la característica o habilidad que está siendo evaluada. En el MRG, que es una extensión del modelo de 2PL que se ha descrito anteriormente, la capacidad del ítem para discriminar entre los niveles del rasgo latente se mantiene constante mientras que la dificultad del ítem se establece en cada "paso del ítem" (Penfield, 2014, p. 10), o cuando una respuesta pasa de una categoría de respuesta a otra. Es decir, si tenemos una escala de respuesta de 4 puntos (p.e., 1, 2, 3, 4) tendríamos *k-1* pasos (parámetros *b*) puesto que la respuesta del sujeto pasa de (a) 1 a 2; (b) 2 a 3, o (c) 3 a 4. Siendo *k* el número de opciones de respuesta al ítem. De este modo, en el modelo tendríamos 3 parámetros *b* o parámetros umbrales.

El modelo se formula en términos de probabilidades acumulativas, de diferencias entre probabilidades acumuladas. La probabilidad de que un sujeto pueda seleccionar una categoría de respuesta *k* o superior viene dada por:

$$P_{ik}^{*}(\theta) = \frac{e^{-1.7a_i(\theta - b_{ik}^{*})}}{1 + e^{-1.7a_i(\theta - b_{ik}^{*})}}$$

En este modelo, hay un parámetro *a* para cada ítem *i* y *k*-1 parámetros de dificultad. Los parámetros b_{ik} nos proporcionan información sobre la probabilidad de cambio de una categoría hacia el siguiente paso o categoría. Además, estos parámetros se encuentran ordenados en modo ascendente, reflejando precisamente este proceso acumulativo.

En el MRG cada ítem viene definido por un parámetro de discriminación (*a*) y *k-1* parámetros umbrales (p.e., si *k*=4 serían b_1, b_2, b_3) (Samejima, 1969). Los ítems con valores de *a* más altos se asume que discriminan de manera más precisa, mientras que los parámetros umbrales corresponden al nivel theta (θ) en el rasgo latente necesario para responder a un cierto punto en la escala (i.e., 1, 2, 3, o 4). En el ejemplo de este artículo, θ (theta) representa la variable latente estimada por el cuestionario (depresión-ansiedad), esta variable viene dada en unidades de desviación típica, es una escala de puntuación tipificada (Media=0 y DT=1).

Directrices para evaluar el comportamiento del ítem

Un paso inicial en el desarrollo o mejora de una escala es llevar a cabo un análisis de los ítems con el objetivo de identificar qué ítems en cada escala están funcionando de una manera apropiada y qué ítems pueden necesitar una revisión o incluso ser sustituidos por otro. Un ítem que funcione bien debe discriminar a los sujetos evaluados en todos los niveles de θ (esto es, debe tener un valor de *a* relativamente alto), utilizar las cuatro opciones de respuesta (cada curva de respuesta debe distribuirse en distintos valores de θ), y reunir información suficiente (es decir un valor relativamente alto información). Este proceso se completa en varios pasos utilizando varias fuentes de información.

En primer lugar, los parámetros *a* pueden ser evaluados siguiendo los criterios de Baker (2001), donde el valor de *a* de .65 se utiliza como el umbral mínimo para que un ítem tenga un funcionamiento aceptable, $a > 1.34$ indica que el ítem tiene un nivel "elevado" de funcionamiento, y $a > 1.69$ indica un ítem con un funcionamiento "muy elevado". Los ítems siempre deben ser evaluados combinando los resultados estadísticos con el contenido de los mismos, es decir, considerando de manera conjunta los aspectos teórico-substantivos de los ítems y sus parámetros estimados. Aquellos ítems que no cumplen con el umbral mínimo ($a < .65$) pueden ser revisados por expertos en el tema. En segundo lugar, se evalúan las CCIs. Los ítems en los que no se ha utilizado alguna de las cuatro opciones de respuesta (es decir, las curvas características se solapan, ver la Figura 2, panel D) y obtienen valores de *a* bajos deben ser revisados por expertos en el tema. En tercer y último lugar, se evalúan las funciones de información para cada uno de los ítems, y aquellos ítems que proporcionan poca información (es decir, aquellos cuya curva información es plana, véase la Figura 3, panel D) también deben ser revisados. Aquellos ítems que superan estas fases se pueden mantener en el test. Por el contrario, aquellos que han sido identificados en el proceso deben ser revisados o sustituidos. Más allá de estos criterios, existen pruebas estadísticas de bondad de ajuste de los ítems y de diagnóstico de los modelos que se utilizan en el proceso de evaluación, que por cuestión de espacio no serán comentadas en este trabajo. Para una buena revisión aplicada de la evaluación del ajuste de ítems en TRI remitimos al lector al excelente trabajo de Ames y Penfield (2015), también se puede acudir a algunas de las referencias más generales que se han ido mencionando a lo largo de este trabajo. Una vez que se adquiere un dominio básico en los modelos, es bastante sencillo captar los detalles adicionales.

Un ejemplo: Escala de Depresión-Ansiedad

Uno de los grandes desafíos de los tribunales de menores y los responsables educativos es encontrar la mejor manera de ayudar a los jóvenes que están experimentando múltiples problemas conductuales y emocionales, y para quienes tanto la asistencia al centro educativo como su rendimiento académico es bajo. Existe una fuerte asociación entre el fracaso escolar y la implicación en el sistema de justicia de menores, especialmente para los jóvenes con problemas relacionados con el abuso de sustancias, trastornos psiquiátricos, y absentismo escolar. Para ayudar a identificar a los estudiantes que tienen cierto riesgo y muestran necesidades, podemos usar varias medidas. En concreto el Washington Assessment of the Risks and Needs of Students (WARNS) es un autoinforme breve (40 ítems, 20 minutos) para estudiantes entre 13 y 18 años de edad. Está diseñado para permitir que los centros educativos, los tribunales y los responsables del sistema educativo puedan evaluar riesgos individuales y necesidades que pueden dar lugar al absentismo escolar y/o el fracaso escolar. Esta medida evalúa experiencias en varios dominios que son críticos para el desarrollo social, emocional y educativo saludable. La WARNS evalúa seis dominios que están relacionados con el absentismo escolar, la delincuencia en la escuela, y otros comportamientos desadaptativos (Hammond, Linton, Smink, & Drew, 2007; Howell, 2003). Las seis escalas incluyen (a) Compromiso con la escuela, (b) Ambiente familiar, (c) Abuso de Sustancias, (d) Desviación del grupo de iguales, (e) Agresión-Desafío, y (f) Depresión-Ansiedad.

Para este ejemplo, nos centraremos en comportamientos o síntomas internos, que un estudiante experimenta comúnmente, pero que pueden ser difíciles de detectar mediante la observación directa en el centro educativo. Por lo tanto, nos vamos a centrar en la escala de Depresión-Ansiedad, que evalúa niveles de comportamientos o síntomas de expresión interna, que tienden a tomar forma como depresión y ansiedad, y puede dar lugar a una intensa tristeza, desesperanza, y problemas de sueño y alimentación (American Psychiatric Association, 2013). La depresión y la ansiedad, cuando son frecuentes o graves, pueden llevar a obstaculizar el funcionamiento físico, social y psicológico. Además, puede dar lugar a comportamientos suicidas, autolesiones, deterioro del funcionamiento cognitivo, y bajo rendimiento escolar (American Psychiatric Association, 2013). El WARNS combina la evaluación de la depresión y la ansiedad dado que en jóvenes estos síntomas ocurren al mismo tiempo (por ejemplo, Hinden et al., 1997). En esta evaluación las puntuaciones fueron diseñadas de manera que niveles más altos de depresión y ansiedad vienen indicados por puntuaciones más altas en la escala. Para favorecer la identificación de aquellos jóvenes con mayor riesgo y necesidades, la mayoría de la información, en el marco de la TRI debe estar en el extremo superior de la distribución del rasgo latente, en nuestro caso en depresión y ansiedad. Por lo tanto, si los ítems de esta escala están funcionando correctamente la función de información del test debería mostrar un pico, proporcionar mayor información, en niveles de Depresión-Ansiedad cercanos al nivel de 2.0 (dos unidades de desviación típica por encima de la media) siendo en estos niveles del rasgo el menor error de estimación. Además, también nos interesa observar las respuestas a cada uno de los ítems de la escala.

Instrumento

La Escala de Depresión-Ansiedad contiene 8 ítems. El formato de respuesta de los mismos se encuentra en una escala de 4 puntos: (a) Nunca, (b) A veces, (c) A menudo, y (d) Siempre que indican el grado en que se producen esos sentimientos. Cuatro de los enunciados pregunta a los jóvenes acerca de síntomas de la depresión. Los otros cuatro ítems abordan los síntomas de ansiedad. Los sentimientos de depresión y ansiedad suelen ser comunes en la adolescencia. Sin embargo, los jóvenes en riesgo informarán sobre estos síntomas con más frecuencia y por lo general por un período prolongado de tiempo. Los cuatro ítems que hacen referencia a la depresión son: (a) *Sentí que nada me podría animar*, (b) *Me sentí hundido, triste, e infeliz*, (c) *Me sentí desesperado sobre el futuro*, y (d) *No me importa nada ni nadie*. Estos ítems evalúan la frecuencia de síntomas tales como tristeza, desesperanza y falta de interés. Los ítems referidos a ansiedad son: (a) *Estaba tan preocupado o alterado por cosas que era difícil concentrarse*, (b) *He tenido problemas para dormir o comer, porque no podía quitarme algo de mi mente (cabeza)*, (c) *Me sentí más tenso, irritado, o preocupado de lo habitual*, y (d) *Me puse tan nervioso, me sentía mal, tenía problemas para respirar, o me sentí tembloroso*. Estos ítems evalúan la frecuencia de síntomas tales como malestar más allá de una experiencia normal, quejas somáticas, tales como náuseas, comer o dormir. La fiabilidad estimada como consistencia interna usando alfa de Cronbach fue de .87. Además, la estructura factorial analizada en estudios previos, ha mostrado que esta escala evalúa un factor dominante (es decir, la depresión-ansiedad), que cumple con el supuesto de la TRI de que la escala sea esencialmente unidimensional.

Participantes

Los participantes fueron adolescentes (*N*=937, 52.8% chicos) entre 11 y 19 años de edad del Estado de Washington en Estados Unidos. El origen étnico de los participantes fue predominantemente caucásicos (50.5%) e hispanos/latinos (32.1%). El 38% de los estudiantes nunca habían sido expulsados del centro de estudios, el 10% con 11 o más expulsiones; 65.5% nunca había sido detenido, el 9.3% con 3 o más detenciones). Los estudiantes completaron la escala bien en la escuela o en una oficina de apoyo local a la educación como parte de la evaluación de riesgo y necesidades.

Resultados

Se estimó el modelo MRG utilizando el paquete de software IRTPRO3, publicado por ssicentral.com. La Teoría de Respuesta al Ítem de resultados informados por los pacientes (*Item Response Theory for Patient-Reported Outcomes*, IRTPRO) es una nueva aplicación para el calibrado de ítems y puntuaciones del test utilizando TRI. IRTPRO3 permite estimar los modelos de 1, 2 y 3PL, el MRG y el Modelo Nominal. Además, dispone de una variedad de métodos de estimación, rutinas para la obtención de puntuaciones, y análisis de ítems incluido el análisis del Funcionamiento Diferencial del Ítem (DIF), siendo un programa que dispone de una interfaz de fácil uso. En el último apartado de este trabajo se ofrece más información acerca del software disponible para el análisis usando la TRI.

La TRI dispone de un estadístico para evaluar la fiabilidad, de manera similar a la fiabilidad como consistencia interna de la TCT. Tal y como se comentó anteriormente la estimación de la fiabilidad según la TCT fue de .87. La estimación de fiabilidad marginal usando la TRI para esta misma escala 8 ítems es de .84. Estos valores se encuentran dentro del mismo rango y llevarían al usuario a la misma conclusión substantiva sobre la estimación observada o latente del rasgo evaluado a partir de estos 8 ítems. Hay que tener en cuenta que a

medida que aumenta la fiabilidad disminuye el error típico de medida, independientemente de trabajar desde la TCT o la TRI. Por lo tanto, cuanto más fiables sean los resultados obtenidos, más precisa será la medición de la variable. Sin embargo, en TRI la fiabilidad es más fácil de entender en términos de la función de información, tal y como se ha comentado en la presentación de la Figura 1. Cuanta mayor información proporciona un test o un ítem, mayor es la precisión en la estimación del rasgo o característica del individuo. Lo más importante en TRI, diferente a la TCT, es que la precisión puede ser evaluada en cada nivel del rasgo psicológico que está siendo medido y por lo tanto puede no ser constante tal y como asume la TCT. La Figura 3, panel A nos puede ayudar a comprender esta idea.

El panel A de la Figura 3 contiene la función de información de la escala de Depresión-Ansiedad (línea continua), y el error típico de medida (línea discontinua). La función de información de la escala se obtiene como la suma de las funciones de información de los ítems en cada nivel del rasgo. Tal y como podemos observar, la escala es más precisa o proporciona más información en torno a los niveles de depresión-ansiedad de 1.5 a 2.0 o por encima de niveles medios de depresión y ansiedad. Por otro lado, el error es mayor, menos precisión, en niveles bajos de la variable (de -1 a -3). Para el propósito de nuestro test, esto puede ser deseable, ya que la finalidad es identificar con precisión aquellos jóvenes con altos niveles de depresión y ansiedad. De hecho, la línea horizontal establecida en la puntuación estimada de 1.50 podría ser el punto de corte para determinar la necesidad de intervención. Si esto fuera así, sería en dicho nivel de la variable donde el instrumento es más preciso.

Tabla 1. Parámetros estimados según el Modelo de Respuesta Graduada para los ítems de Escala de Depresión de la WARNS.

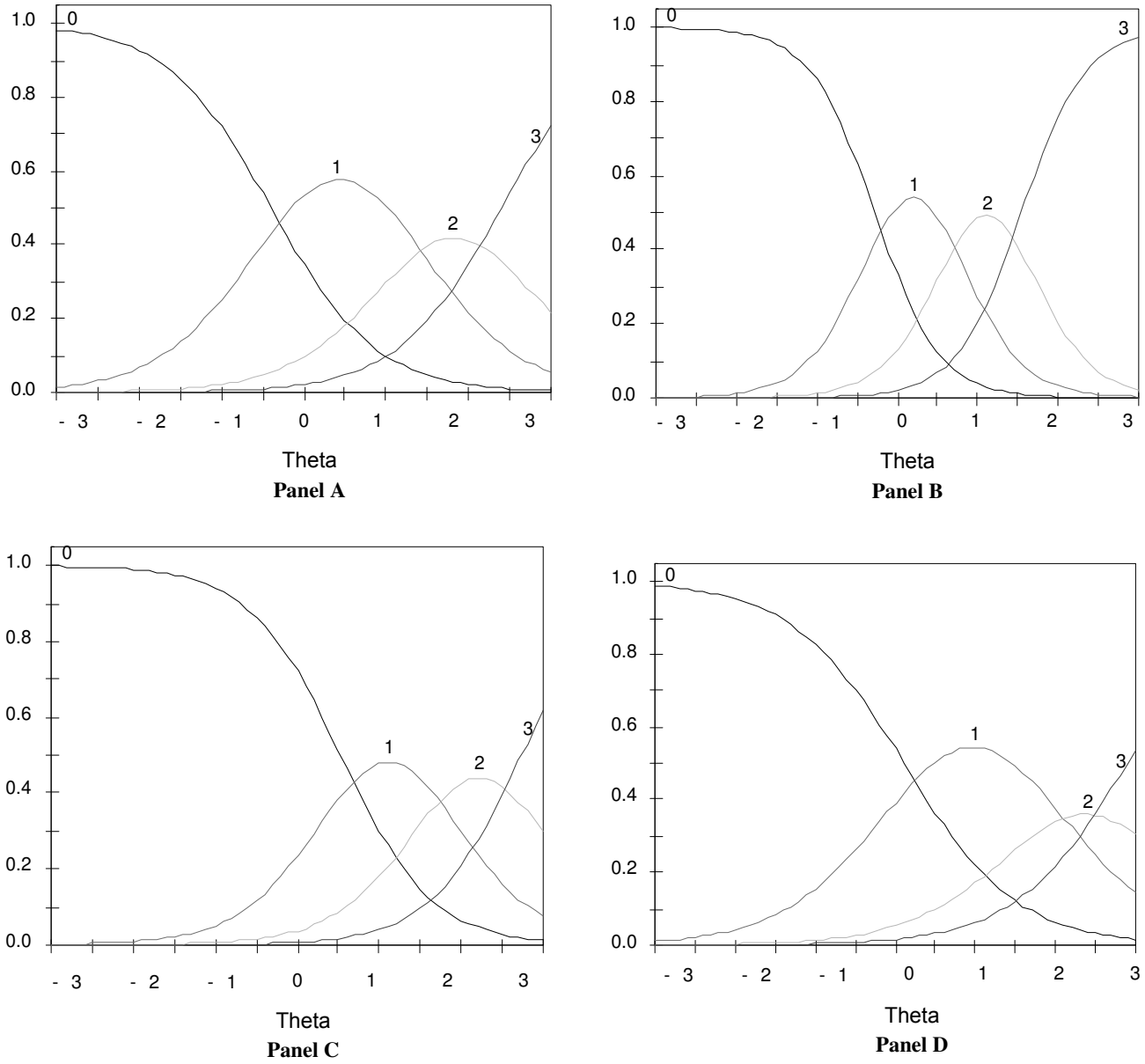
Item	*a*	*s.e.*	b_1	*s.e.*	b_2	*s.e.*	b_3	*s.e.*	$S\text{-}\chi^2$	*p*
1	0.51	0.08	1.65	0.27	2.98	0.46	3.92	0.60	54.87	.40
2	1.58	0.11	-0.40	0.06	1.27	0.08	2.40	0.15	50.58	.22
3	2.47	0.16	-0.46	0.05	0.77	0.05	1.65	0.08	84.37	<.01
4	2.51	0.17	-0.28	0.05	0.68	0.05	1.54	0.08	52.67	.08
5	1.82	0.12	-0.26	0.06	1.28	0.08	2.46	0.14	56.34	.10
6	2.79	0.19	-0.14	0.05	1.02	0.06	2.07	0.10	42.96	.23
7	1.81	0.14	0.53	0.06	1.69	0.10	2.73	0.17	60.08	.08
8	1.42	0.11	0.11	0.06	1.83	0.12	2.90	0.20	58.42	.22

a = parámetro de discriminación; s.e. = error típico para la estimación obtenida; $b_{1\text{-}3}$ *= parámetros umbrales o valores de dificultad que reflejan el cambio de una categoría de respuesta a otra.*

En la Tabla 1 se presentan los parámetros estimados junto a sus errores típicos para cada uno de los 8 ítems. Hay que tener en cuenta que hay tres estimaciones para el parámetro *b*, una para cada umbral dado que tenemos 4 opciones de respuesta, es decir, *k-1* umbrales. La segunda columna de la tabla contiene el parámetro *a* (discriminación del ítem), según los criterios anteriormente comentados, todos los ítems presentan valores de discriminación buenos, siendo el ítem 6 el más discriminativo y el ítem 1 el menos discriminativo con un valor ligeramente inferior al criterio establecido. Las columnas 4, 6 y 8 contienen los parámetros umbrales referidos a la habilidad mínima para pasar de una respuesta de 1 a 2, 2 a 3, y de 3 a 4, respectivamente. Tal y como podemos observar en la tabla, según los valores de b_1, no es necesario tener un nivel muy alto de depresión-ansiedad (por ejemplo, <0 para 5 de los 8 ítems) para pasar de una respuesta de 1 (*nunca*) a una de 2 (*a veces*). Sin embargo, si es necesario tener un nivel alto en el rasgo (por ejemplo, > 2.0 en 6 de los 8 ítems) para pasar de una respuesta de 3 (*a menudo*) a una de 4 (*siempre*). Para esta escala esto es lo apropiado, ya que para elegir la opción de respuesta 4 en comparación con la 1 ó 2 se requiere un nivel mucho más alto de depresión-ansiedad, puesto que esta escala es más precisa y proporciona más información en niveles altos del rasgo. También podemos observar que los errores típicos de estimación son más altos en los parámetros de los umbrales más altos en comparación al resto, este resultado posiblemente está relacionado con la menor proporción de sujetos que se sitúan en esas categorías de respuesta alta, tal y como es de esperar. Además, podemos observar que los valores de *b*, en cada ítem, están ordenados de modo ascendente, aumentan de valor de b_1 a b_3. Esta cuestión es importante, ya que se requiere de niveles más altos de depresión-ansiedad para dar una respuesta de 4 comparada con una respuesta de 3 o una de 2 o una de 1. Las dos últimas columnas de la tabla informan de un valor de chi-cuadrado y de su probabilidad asociada (*p*), que hacen referencia a una prueba de ajuste estadístico del ítem al MRG (Orlando & Thissen, 2000, 2003). En este caso, sólo un ítem, ítem 3, tiene un valor estadísticamente significativo por debajo de .01. Después de tener en cuenta el ajuste por los múltiples contrastes estadísticos, donde debemos considerar un nivel de significación más bajo, y de revisar el comportamiento del ítem, no parece existir un problema con este ítem.

La Figura 2 representa las curvas características del ítem (CCI) de cuatro de los ocho ítems de la escala (ítems 2, 4, 7 y 8). Estos gráficos son quizás una de las principales ventajas del análisis de ítems y desarrollo de un test bajo la TRI. A través de estas CCIs se puede fácil y rápidamente tener una visión del funcionamiento de los ítems a través del continuo de habilidad y para cada una de las opciones de respuesta al ítem. El ítem 2 en el Panel A (*Me sentí hundido, triste, e infeliz*), por ejemplo, se enfoca en el centro de la distribución, con un buen uso de todas las alternativas de respuesta y proporciona buena información teniendo en cuenta el apuntamiento de las curvas asociado con valores altos de *a* (1.58). El ítem 4 (*No me importa nada ni nadie*), en el Panel B, sigue una tendencia similar. Por el contrario,

Figura 2. Panel A: Curvas Características del Ítem para el ítem 2; Panel B: Curvas Características del Ítem para el ítem 4; Panel C: Curvas Características del Ítem para el ítem 7; y Panel D: Curvas Características del Ítem para el ítem 8.

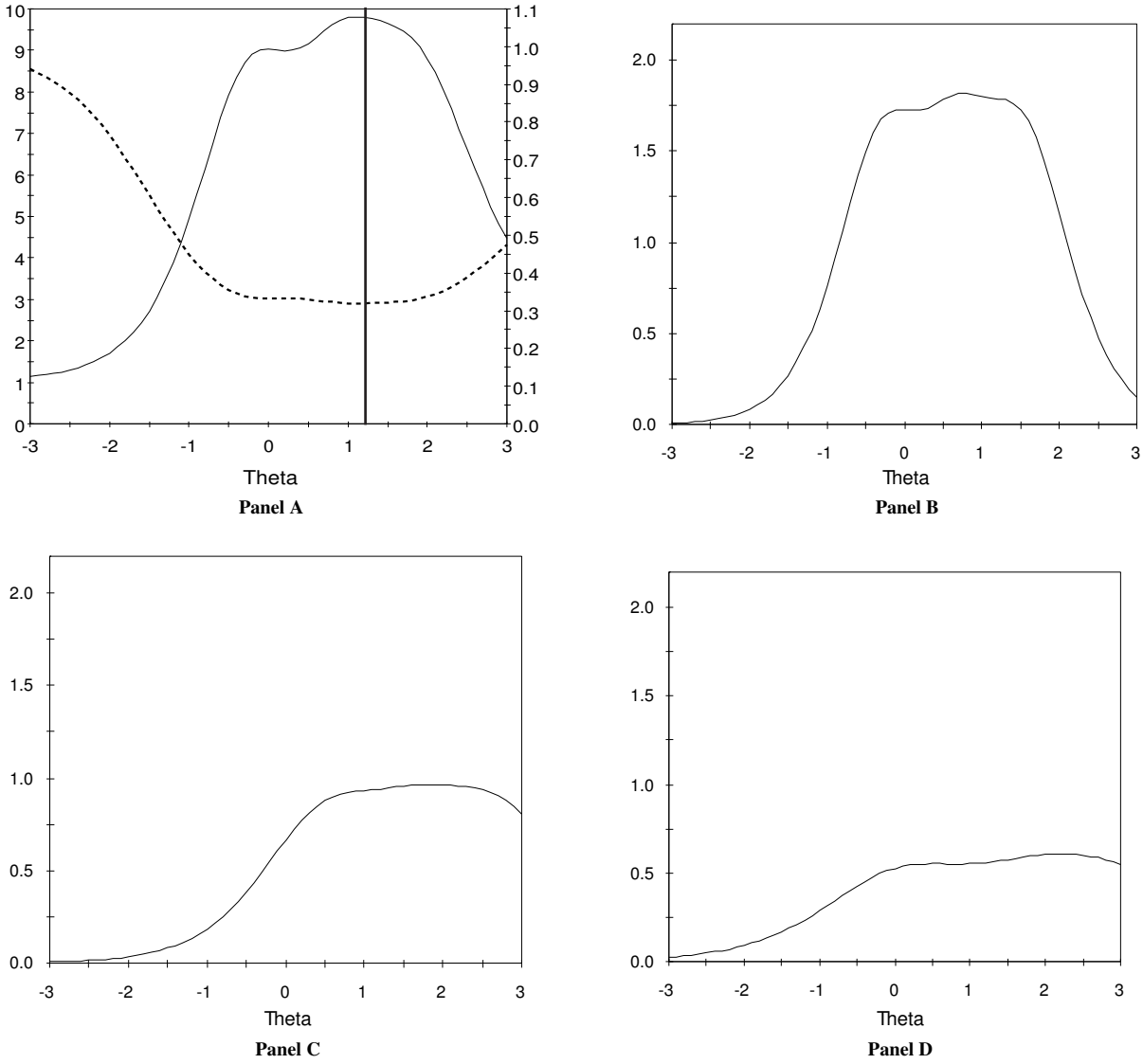


el ítem 8 (*Me puse tan nervioso, me sentía mal, tenía problemas para respirar, o me sentí tembloroso*) en el Panel D tiene curvas claramente desplazadas hacia el extremo superior de la distribución del rasgo, con un menor apuntamiento, que indican un valor menor de *a* (1.42). Esto es coherente con el contenido de este ítem, que es más específico y sobre emociones más extremas. El ítem 7, ver en el panel C, sigue la misma tendencia.

En la Figura 3 podemos observar las diferencias en la cantidad de información que aportan estos ítems y su distribución a lo largo del continuo del rasgo. El panel B muestra que el ítem 4 proporciona la mayoría de la información, en comparación con el resto de ítems mostrados en la Figura 3, y que la curva se encuentra ligeramente centrada por encima del nivel medio en Depresión-Ansiedad. En comparación, los ítems 7 y 8, Panel C y D respectivamente, presentan una función de información más achatadas y desplazadas al extremo superior de la distribución del rasgo. A pesar de que estos apuntamientos no son tan elevados como el del ítem 4, y sugiere una posible sustitución o revisión, estos ítems muestran tres características importantes que son valiosas para la escala. En primer lugar, todas las opciones de respuesta han sido respondidas y los valores de *a* son aceptables. En segundo lugar, se orientan a niveles altos del rasgo (por encima del promedio), donde la calidad de los ítems es necesaria. En tercer lugar, y quizás lo más importante, su contenido está alineado con el constructo que se está midiendo siendo un contenido esencial y representativo del mismo. Ahora, animamos al lector a tratar de imaginar todas las curvas de información de los ítems solapadas entre sí para formar la Figura del panel A, es decir, la función de información del test. La función de información de la escala de depresión-ansiedad se obtiene como la suma de las funciones de información de los ítems.

En consecuencia, considerando las estimaciones de los parámetros, las CCIs, las funciones de información de los ítems, la función de información del test, y la revisión substantiva de los ítems, todo en conjunto sugiere que la escala de depresión-ansiedad funciona según lo previsto. El análisis de ítems basado en la TRI nos ha permitido esta comprobación visual sobre la escala y nos ha proporcionado un componente de evidencia, en principio, para construir

Figura 3. Panel A: Función de Información del Test para la escala de Depresión (línea continua) y Error típico de medida (línea discontinua); Panel B: Función de Información para el ítem 4; Panel C: Función de Información para el ítem 7; y Panel D: Función de Información para el ítem 8.



un argumento de validez para el uso de estas puntuaciones en la toma de decisiones sobre los estudiantes con respecto a los niveles de depresión y ansiedad.

Teoría de Respuesta al Ítem: Software disponible

Hay muchas, quizás demasiadas, opciones de software para realizar análisis psicométricos, incluyendo los análisis basados en TRI. Debido a las limitaciones de espacio, es imposible proporcionar todas las ventajas y desventajas de cada programa. De hecho, sería posible escribir un extenso capítulo solo sobre este tema. Dicho esto, algunos de los principales programas de ordenador utilizados son los paquetes estadísticos de uso general (por ejemplo, Mplus, SAS, R) que también pueden estimar modelos de TRI. También hay una gran cantidad de paquetes de software que se han construido específicamente para el análisis de TRI generales (por ejemplo, BILOG-MG3, FlexMIRT, WINMIRA), o incluso para aplicaciones más específicas tales como los resultados basados en pacientes (por ejemplo, IRTPRO), elaboración de tests (por ejemplo, PARSCALE), y para análisis más concretos, tales como el análisis del sesgo de los ítems a través del análisis de funcionamiento diferencial del ítem (por ejemplo, IRTLRDIF). Algunos de estos paquetes son de libre acceso (por ejemplo, IRTLRDIF; R packages), otros tienen una versión para estudiantes (por ejemplo, IRTPRO 3) para experimentar con ellos o ejecutar análisis de TRI sencillos, mientras que otros paquetes requieren comprarlos para uso. Para comenzar es posible explorar estas diferentes opciones en sitios web tales como www.ssicentral.com, un importante distribuidor de este tipo de software. Como se puede intuir, el extenso número de opciones puede llegar a ser abrumador. Además de software, hay libros de texto que nos pueden ayudar a entender la forma de realizar el análisis psicométrico con programas tanto de TRI especifícos como programas estadísticos generales (por ejemplo, De Ayala, 2009; Finch, Immekus, & French, 2014).

Artículo recibido: 23/04/2016
Aceptado: 09/06/2016

Conflicto de intereses

Los autores de este trabajo declaran que no tienen conflicto de intereses.

Referencias

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (6th ed.). Washington, DC: Author.

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice, 34,* 39-48. doi:10.111/emip.12067

Baker, F. B. (2001). *The basics of item response theory*. Recuperado de http://files.eric.ed.gov/fulltext/ED458219.pdf

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO [Computer software]*. Chicago, IL : Scientific Software.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York, NY: Guilford Publications.

Embretson, S. E., & Reise, S. P. (2009). *Item response theory for psychologists*. New York, NY: Psychology Press.

Finch, W. H., French, B. F., & Immekus, J. C. (2014). *Applied psychometrics using SAS.* Charlotte, NC: Information Age Publishing.

International Test Commission (2001). International Guidelines for Test Use. *International Journal of Testing, 1*(2), 93-114. doi:10.1207/S15327574IJT0102_1

Hambleton, R. K., Swaminathan, J., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* New Bury Park, CA: SAGE Publications, Inc.

Hammond, C., Linton, D., Smink, J., & Drew, S. (2007*). Dropout Risk Factors and Exemplary Programs.* Clemson, SC: National Dropout Prevention Center, Communities In Schools Inc.

Hinden, B. R., Compas, B. E., Howell, D. C., & Achenbach, T. M. (1997). Covariation of the Anxious-Depressed Syndrome During Adolescence: Separating Fact From Artifact. *Journal of Consulting and Clinical Psychology, 65*, 6-14. doi:10.1037/0022-006X.65.1.6

Howell, J. C. (2003). *Preventing and reducing juvenile delinquency: A comprehensive framework*. Thousand Oaks, CA: Sage.

López-Pina, J. A. (1995). *Teoría de Respuesta al Ítem: Fundamentos*. Barcelona: PPU.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mellenbergh, G. J. (1994). Generalized Linear Item Response Theory. *Psychological Bulletin, 115*, 300-307. doi:10.1037/0033-2909.115.2.300

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los ítems. *Papeles del Psicólogo, 31*(1), 57-66. Recuperado de www.papelesdelpsicologo.es/pdf/1796.pdf

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Methods.* Chicago, IL: Irwin.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64. doi:10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of the S-X^{2}: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27,* 289-298. doi:10.1177/0146621603027004004

Penfield, R. D. (2014). An NCME Instructional Module on Polytomous Item Response Theory Models. *Educational Measurement: Issues and Practice, 33*, 36-48. doi:10.111/emip.12023

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Recuperado de http://www.psychometrika.org/journal/online/MN17.pdf

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer Verlag.