La escasa aplicación de la teoría de respuesta al ítem en tests de ejecución típica

The limited application of the Item Response Theory to typical performance tests

FACUNDO JUAN PABLO ABAL
GABRIELA SUSANA LOZZIA
MARÍA ESTER AGUERRI
MARÍA SILVIA GALIBERT
HORACIO FÉLIX ATTORRESI*

Universidad de Buenos Aires, Argentina

Resumen

El objetivo de este trabajo fue exponer los factores que explican la escasa aplicación de la Teoría de Respuesta al Ítem (TRI) en tests de ejecución típica. Una revisión exploratoria de la literatura psicométrica permitió identificar dos factores: el histórico, que remite a problemas en la difusión de la TRI por fuera del campo educativo y a las demandas sociales que privilegiaron su aplicación en el contexto educativo; y el factor asociado a la aplicación, que describe los obstáculos que en la actualidad dificultan la puesta en práctica de los modelos en tests de ejecución típica. Aunque ambos factores fueron relativamente superados, el impacto de estos en la producción científica del área es aún significativo.

Palabras clave: medición de actitudes, medición de la personalidad, psicometría, Teoría de Respuesta al Ítem, test de ejecución típica.

Abstract

This study was aimed to evidence the factors behind the limited application of the Item Response Theory (IRT) to typical performance tests. An exploratory review of psychometric literature has led to identify two factors. The historical factor refers to the problems of spreading the IRT outside the educational field and to the social demands that favored its application in the educational context. And the application-related factor refers to the theoretical and operating obstacles which currently hinder the implementation of models in typical performance tests. Although both factors can be considered as overcome, their impact on the scientific production in this field is still significant.

Keywords: attitude measurement, Item Response Theory, personality measurement, pyschometrics, typical performance test.

ARTÍCULO DE REVISIÓN
RECIBIDO: 31 DE AGOSTO DEL 2009 - ACEPTADO: 27 DE ABRIL DEL 2010

* Correspondencia: Zuviría 5691 (CP1439) Ciudad Autónoma de Buenos Aires. E-mail primer autor: fabal@psi.uba.ar

La teoría de Respuesta al Ítem (TRI) es un enfoque psicométrico introducido en la década de los cincuenta con el fin de superar algunas limitaciones teóricas de la Teoría Clásica de Tests (TCT). Su característica principal es que toma a cada ítem como unidad de análisis vinculando el nivel de rasgo que posee un individuo con la probabilidad de dar la respuesta clave al reactivo (Baker, 2001; Hambleton & Swaminathan, 1985; Lord, 1980).

El objetivo sustancial de la TRI es la construcción de instrumentos de medición con propiedades invariantes entre poblaciones, lo que Wright (1968) definió como *tests libres de muestra* (*sample-free test*). Asimismo, la TRI permite mediciones invariantes más allá de la composición del instrumento. Esto es, si se varía el conjunto de ítems utilizado para medir el nivel del rasgo en un individuo, la puntuación estimada no se altera de forma significativa, aunque eventualmente hayan cambiado las propiedades psicométricas de los reactivos (Hambleton & Swaminathan, 1985).

Desde hace varios años se observa un incremento progresivo de la aplicación de los modelos de la TRI al análisis de ítems pertenecientes a tests de ejecución típica. Según Cronbach (1972), este tipo de tests se caracteriza por plantear al evaluado un conjunto de situaciones de la vida cotidiana con el fin de obtener información respecto de cómo se comporta usualmente. En otras palabras, los tests de ejecución típica indagan un área no-cognitiva de la conducta humana como son los rasgos de la personalidad, intereses y actitudes (Martínez-Arias, 1995). En contraposición, las pruebas que plantean problemas para que el individuo muestre su capacidad de resolución se denominan tests de ejecución máxima. Estas últimas tienen como objetivo evaluar habilidades, aptitudes y rendimiento.

Existen investigaciones que recientemente han confirmado la utilidad de estudiar la respuesta de los individuos a los ítems de ejecución típica a partir de las premisas de la TRI. En función del tipo de respuesta que admite el ítem, se han efectuado aplicaciones para puntuaciones dicotómicas (p. e., Cadavid, Delgado & Prieto, 2007; Childs, Dahlstrom, Kemp & Panter, 2000; Ferrando & Chico, 2000; Reise & Waller; 1990) y politómicas (p. e., Gomez & Fisher, 2005; Gray-Little, Williams & Hancock, 1997). Pero los beneficios de la TRI en el contexto de la medición de variables de personalidad, actitudes e intereses no están limitados exclusivamente al análisis de las propiedades psicométricas de los reactivos. Su aplicación puede extenderse con el objetivo de encontrar nuevas soluciones para la construcción de tests adaptativos informatizados (p. e., Aguado, Rubio, Hontangas & Hernández, 2005; Reise & Henson, 2000), el estudio del funcionamiento diferencial de los ítems (p. e., Escorial & Navas, 2007; Smith, 2002), la determinación de la cantidad óptima de opciones de respuesta (p. e., Hernández, Espejo & González-Romá, 2006; López-Pina, 2005) y la detección de patrones de respuesta anómalos (Reise & Waller, 1993; Zickar & Drasgow, 1996).

A pesar de los avances registrados en la actualidad, a lo largo de la historia de la TRI se puede observar cierta desatención de los teóricos que utilizan tests de ejecución típica sobre estos procedimientos de la psicometría moderna. Damarin (1970) advirtió sobre esta problemática muy tempranamente. Tiempo después, Thissen, Steinberg, Pyszczynski y Greenberg (1983) resaltaron que el análisis de tests de ejecución típica con TRI estaba largamente retrasado respecto de los avances que se habían alcanzado hasta aquel momento con variables de rendimiento y habilidades. También durante la década de los ochenta, Kline (1983) y Roskam (1985) señalaron en sus libros la necesidad de poner más empeño en el estudio de los potenciales beneficios de la TRI en el dominio de los tests de ejecución típica. Pero estas demandas no tuvieron una repercusión importante entre los investigadores del área.

Hacia 1994, Ferrando aseguró que una revisión general de la literatura acerca de la TRI seguía mostrando muy poca o nula aplicación de

sus modelos al estudio de la personalidad. Según este autor, al efectuar búsquedas específicas solo se encontraban algunos artículos aislados. Es justamente en la década de los noventa cuando Reise, Waller y sus colaboradores publicaron una serie de artículos que instalaron el tema entre los especialistas como un campo de investigación promisorio (p. e., Reise, 1999; Reise & Waller, 1990; Waller, Tellegen, McDonald & Lykken, 1996).

Una revisión más actual de la bibliografía psicométrica sigue dejando en evidencia la notoria infrecuencia con que se aplicó la TRI para analizar ítems de ejecución típica (Morizot, Ainsworth & Reise, 2007). Reise y Henson (2003) resaltaron que el escaso uso de la TRI en la medición de la personalidad es indiscutible al compararlo con el enorme progreso que ha tenido la aplicación en ítems de habilidades. La magnitud de esta desproporción es tal que si se observa la evolución de la TRI podría afirmarse que fue utilizada casi exclusivamente para la modelización de tests de rendimiento, habilidades y aptitudes (Embretson & Reise, 2000; Rojas & Pérez, 2001).

Si la TRI ha demostrado ser una teoría potente que supera limitaciones teóricas y prácticas de la TCT, ¿cuáles son las razones, a más de cincuenta años de sus primeras formulaciones, de que haya tenido tan baja aplicación para la modelización de constructos propios del área no-cognitiva de la conducta humana? El objetivo de este trabajo es exponer los factores mencionados en la literatura psicométrica como causantes de esta escasa utilización de la TRI en tests de ejecución típica y que, en alguna medida, también pueden justificar esta desproporción en las aplicaciones en función del tipo de constructo.

Parece difícil explicar mediante una causa única la dispar aplicación de los modelos de la TRI en función del tipo de constructo. Más complejo aún resulta revelar por qué la TRI ha sido relativamente poco utilizada en tests de ejecución típica. Se ha intentado explicar este fenómeno hallando sus causas en diferentes esferas tanto teóricas como prácticas. Una revisión exploratoria de la bibliografía psicométrica ha permitido identificar dos factores: factores históricos y de aplicación. Mientras los primeros justifican el escaso uso en tests de ejecución típica remitiendo al surgimiento de la TRI, los segundos explican cuáles son los obstáculos actuales que dificultan las aplicaciones. A continuación se desarrollarán las características específicas de ambos factores.

Factor histórico

Problemas asociados al surgimiento y difusión de la TRI

El desarrollo formal de la TRI se produjo principalmente en el contexto de la medición educativa. Los primeros modelos de la TRI fueron propuestos por Lord (1952) y Rasch (1960), entre otros. Estos autores utilizaron los modelos basados en la distribución de la ojiva normal y logística para el análisis de ítems de rendimiento y aptitudes puntuados dicotómicamente. Jones y Thissen (2007) también reconocieron la influencia de una corriente sociológica de corte cuantitativo en los albores de la TRI, pero su contribución fue mínima y embrionaria comparada con los posteriores aportes generados por los especialistas en educación.

La escasa aplicación de la TRI a tests de ejecución típica se debe en gran medida a este origen educativo del que emergieron los modelos (Chernyshenko, Stark, Chan, Drasgow & Williams, 2001; Martínez-Arias, 1999; Morizot et al., 2007; Reise, 1999; Rojas & Pérez, 2001). Reise (1999) explicó este fenómeno como un retraso temporal de la aplicación de la TRI a otras variables ajenas al campo de la educación.

Los modelos se fueron perfeccionando solo a partir de obstáculos surgidos en elaboración de pruebas de rendimiento, con lo cual, además, se observó una postergada incorporación de herramientas de análisis propias de otras disciplinas. Un ejemplo de este fenómeno puede encontrarse incluso en la modelización con TRI de tests de ejecución máxima. Desde

sus orígenes, la TRI se aplicó en el contexto educativo a la medición de habilidades y aptitudes al margen de los potenciales aportes que podían brindarse desde la psicología cognitiva; esta integración, según Prieto y Delgado (1999), no se produce sino hasta hace pocos años.

Durante los primeros años de la historia de la TRI, los teóricos estuvieron concentrados en el desarrollo y en la complejización de modelos para ítems dicotómicos. Este formato de respuesta resultaba ideal para las aplicaciones educativas, puesto que era y es el más frecuentemente utilizado en tests de ejecución máxima que puntúan con 1 el acierto y con 0 el fallo. La limitación que presentaban estos modelos dicotómicos de la primera generación de la TRI era la dificultad de dar un tratamiento satisfactorio a ítems puntuados politómicamente, formato por excelencia de ítems de tests de ejecución típica. Para Reise (1999), esta limitada aplicabilidad de los modelos a ítems dicotómicos fue clave para que no se despertara el interés de los investigadores que provenían del campo de la psicología de la salud y la personalidad.

Hacia 1969, Samejima describió el primer modelo para ítems de respuesta politómica graduada. Aunque su autora lo propuso inicialmente para ítems de habilidades con respuestas parcialmente correctas, las características del modelo permitieron también extender su aplicación al análisis de reactivos con escalas tipo Likert. Si bien esto inauguró una nueva generación de desarrollos en el área psicométrica, la revisión efectuada para esta investigación no mostró un aumento importante de las aplicaciones en tests de ejecución típica en este periodo.

Sin embargo, la inexistencia de modelos politómicos no puede explicar por sí sola la baja aplicación de la TRI a tests de ejecución típica. Como señaló Hambleton (1997), antes de la aparición de los modelos para datos politómicos, los modelos dicotómicos cubrían todas las necesidades. Según este autor, era habitual dicotomizar los datos politómicos para poder analizar los ítems con modelos dicotómicos. Por otra parte, no todos los tests de ejecución típica usados en esa época presentaban un formato de respuesta politómica. Algunos inventarios (p. e., MMPI) estaban construidos con ítems de dos opciones de respuesta, por lo que no hubiese requerido la dicotomización de las respuestas.

Otro factor histórico que explica la escasa aplicación de la TRI en tests de ejecución típica puede encontrarse en la pobre difusión que se hizo acerca de su potencial para evaluar la personalidad, las actitudes y los intereses. Para Childs et al. (2000), la baja utilización se explica por el desconocimiento de los beneficios que conlleva el uso de esta teoría. Si los especialistas no conocen las ventajas de la TRI por sobre la TCT no intentarán su aplicación. Y, como señalaron Lange y Honran (1999), salvo por algunas excepciones, el potencial de esta teoría en la medición con tests de ejecución típica fue largamente ignorado.

Es innegable que la escasa difusión de la TRI también se asocia a su nacimiento en el campo educativo. Al ser desarrollados en un contexto donde abundan los tests de ejecución máxima, la descripción de los modelos adoptó nombres asociados a las pruebas que evalúan rendimiento para interpretar características matemáticas que son esencialmente neutrales frente a cualquier tipo de constructo (Ositini & Nering, 2005; Zickar & Ury, 2002). Este problema afecta incluso actualmente la difusión de la TRI por fuera de la medición educativa. En efecto, gran parte de la bibliografía sobre TRI sigue explicando los modelos apelando a estas denominaciones en lugar de recurrir a nombres neutros, lo que no contribuye a motivar el interés de investigadores de otros contextos.

A modo de ejemplo, la curva característica del ítem (CCI), un concepto básico de la TRI, se define frecuentemente en la literatura como la representación de una función que describe la probabilidad de contestar de forma correcta a un ítem para todo nivel de habilidad. Los intentos de definir de manera neutra la CCI se refieren a la

probabilidad de dar una respuesta determinada al ítem (opción-clave) en función del nivel del rasgo. En los ítems de rendimiento máximo, la clave es la respuesta correcta, y en los de ejecución típica, es aquella opción que indica la presencia de un nivel mayor de rasgo en el individuo. Nunnally y Bernstein (1995) llamaron *respuesta alfa* a la opción-clave, mientras que Ositini y Nering (2005) la denominaron *opción positiva*.

Como es posible apreciar, también es cierto que resulta difícil eliminar la impronta educativa en la definición de la CCI si se desea pasar a una terminología neutra. La definición neutra de la CCI requiere de una explicación más extensa y detallada que la ligada al campo educativo. Según Ostini y Nering (2005), la popularidad de los términos derivados de los test de ejecución máxima para describir los modelos de la TRI se debe a que estos ayudan a entenderlos más intuitivamente. Tanto la exposición didáctica como la comprensión de estos conceptos se ven facilitadas al asociarlos a nociones de amplia experiencia colectiva como son los exámenes de las instituciones educativas.

Problemas vinculados a la demanda social

Calero y Padilla (2007) señalaron otro factor que ha incidido en un momento histórico de la evolución de la TRI. Según estos autores, los primeros intentos de utilizarla para el análisis de tests de ejecución típica, especialmente en la medición de las actitudes, comenzaron durante el desarrollo y expansión de la TRI como teoría psicométrica

A pesar de las experiencias pioneras, el progreso de esta línea siempre fue más lento que el que se produjo con las variables de habilidad. Calero y Padilla (2007) atribuyen esto a que durante varios años los investigadores con conocimiento sobre TRI desplazaron su interés hacia las pruebas de rendimiento máximo. La razón de la preferencia fue porque estas presentaban una mayor exigencia social por sus objetivos de certificación académica. A causa de un escrutinio público constante y una fuerte demanda legal de las minorías, la medición en el campo educativo estuvo obligada a revisar y perfeccionar sus prácticas. En contraste, la evaluación con tests de ejecución típica ha tenido pocas presiones legales que propicien un replanteo de los procedimientos utilizados. Sin un apremio judicial, se retrasó el tratamiento de las posibles innovaciones (Reise & Henson, 2003).

Durante los años setenta se recuperó la atención de los expertos sobre la aplicación de la TRI a tests de ejecución típica (p. e., Andrich, 1978; Bejar, 1977; Kuncel, 1973). Desde entonces, su empleo aumentó de forma lenta pero sostenida extendiéndose al campo de la evaluación de la personalidad normal y clínica (p. e., Carter & Wilkinson, 1984; Hendryx, Haviland, Gibbons & Clark, 1992; Ozer & Reise, 1994; Schaeffer, 1988), la epidemiología psiquiátrica (Reiser, 1989) y el mercadeo (Singh, Howell & Rhoads, 1990), entre otros contextos.

En resumen, la escasa aplicación de la TRI en el área de la medición con tests de ejecución típica podría explicarse desde un plano histórico, en parte, porque una importante cantidad de investigadores desconocieron durante años sus ventajas (problemas asociados al surgimiento y difusión de la TRI) y, por otra parte, porque aquellos expertos que la conocían se volcaron mayoritariamente a buscar soluciones a las problemáticas de la medición educativa (problemas vinculados a la demanda social).

Factor asociado a la aplicación de la TRI en la actualidad

Las investigaciones aplicadas con tests de ejecución típica realizadas en los últimos años han permitido a los especialistas identificar las limitaciones que puede tener la TRI al intentar modelizarlos. Aunque suponen que tendrá hacia el futuro un importante rol en el análisis de tests de ejecución típica, los autores más optimistas reconocen que el uso de la TRI tiene barreras técnicas que limitan su aplicación rutinaria (Reise &

Henson, 2003; Steinberg & Thissen, 1995). Los más pesimistas consideran que son justamente estos obstáculos los que hacen que la TRI difícilmente pueda reemplazar a la TCT en la medición con tests de ejecución típica (p. e., Aiken, 1999).

Una de las dificultades de aplicación más mencionadas es el cumplimiento del supuesto de *unidimensionalidad* del rasgo latente requerido por los modelos de la TRI más utilizados (Aiken, 1999; Morizot et al., 2007). Los constructos medidos con tests de ejecución típica suelen ser más complejos que las variables que habitualmente se miden con tests de rendimiento máximo. Como aseguraron Fraley, Waller y Brennan (2000) refiriéndose a la medición de la personalidad, los esfuerzos actuales se orientan a la evaluación de dimensiones aisladas aunque muchas de estas variables son inherentemente multidimensionales. No es posible emplear un modelo explicativo unidimensional para aplicarlo a un constructo de naturaleza multidimensional sin que se produzca una representación inaceptable tanto de las propiedades de los ítems como de las personas evaluadas. Como intento de solución para este problema se utilizan técnicas de agrupamiento, como el análisis factorial, para analizar los datos de los tests y aislar distintos subgrupos de ítems que presentan un comportamiento similar. Cada uno de estos subgrupos es tratado por un subtest distinto que representa a cada una de las dimensiones del constructo medido, lo que lleva a perder la información interrelacional entre los subtests (Hambleton, 1997). En definitiva, el requerimiento de unidimensionalidad de la TRI obliga al investigador a realizar una definición conceptual de la variable que recorta las interrelaciones existentes en el marco de una teoría. Reise (1999) reconoció que esta es una de las desventajas más importantes que tiene la TRI al ser aplicada en tests de ejecución típica.

La complejidad teórica de los constructos medidos con tests de ejecución típica también se traduce en dificultades para su operacionalización. Aiken (1999) destacó que el mayor grado de ambigüedad de los ítems de personalidad respecto de los ítems de rendimiento juega en contra al pretender alcanzar la unidimensionalidad requerida por los modelos de la TRI.

Otras dificultades surgen frente a la aplicación de modelos de la TRI a tests diseñados y validados a partir de la TCT. Como aseguraron Embretson y Reise (2000) la TRI ha modificado varias de las reglas clásicas que posibilitaban la medición en psicología. Para Steinberg y Thissen (1995), algunas de estas nuevas reglas pueden restringir el uso de la TRI en la medición con tests de ejecución típica. Desde la perspectiva clásica, este tipo de tests suele utilizar una gran cantidad de reactivos que, en general, son poco consistentes. Esto se deriva de que para la TCT los tests más largos son más fiables que los tests cortos. Por el contrario, la TRI resulta especialmente adecuada para mediciones con un conjunto reducido de ítems altamente consistentes.

A diferencia de la TCT, que requiere una mayor cantidad de ítems que discriminen en los valores medios del rasgo, la TRI necesita que los reactivos elaborados cubran todo el espectro del constructo. Esto significa que se deben redactar ítems con contenidos que discriminen en todos los niveles del rasgo incluidos los extremos. En la medición de una habilidad puede resultar relativamente más fácil predecir el nivel de dificultad que tendrá un ítem y así anticipar su capacidad para discriminar en un nivel particular del rasgo latente. Sin embargo, la experiencia de Reise y sus colegas en la construcción de ítems para inventarios de personalidad mostró que esta tarea se torna compleja (Flannery, Reise & Widaman, 1995; Reise & Waller, 1990).

Si bien la TCT consiente la construcción de ítems redundantes para aumentar la consistencia interna de la escala, desde la perspectiva de la TRI se encuentra un límite con el supuesto de *independencia local*. Este requisito de los modelos de la TRI exige que, para cada nivel del rasgo medido, las respuestas a distintos ítems deben ser estadísticamente independientes. Como aseguraron

Steinberg y Thissen (1995, 1996), la violación de este supuesto puede aparecer en tests de ejecución típica si se utilizan ítems con enunciados similares o exactamente opuestos (utilizados en ocasiones desde TCT también para examinar si el evaluado responde de forma consistente).

Los intentos por sortear estos problemas llegaron cuando algunos autores decidieron adoptar una postura pragmática (Ferrando, 1994; Ferrando & Chico, 2000; Reise & Waller, 1990). El objetivo común de estos estudios consistió en considerar como exploratorios los intentos de aplicación. De esta manera, la pretensión fue evaluar hasta qué punto algunos modelos, los cuales habían sido diseñados para la medición de rendimiento máximo, podían ser adecuados en tests de ejecución típica.

Como se ha podido observar, las limitaciones para aplicar la TRI a datos obtenidos con tests de ejecución típica surgen de contraponer las particularidades de los constructos medidos por estos con las que caracterizan a las variables medidas con tests de rendimiento máximo. Esto también confirma el peso preponderante que tiene el factor histórico. La comparación entre estos tipos de constructos es útil dado que los modelos de la TRI se basan en una descripción racional de la forma en que un individuo contesta al ítem a partir de la cual se presupone una curva de probabilidad de respuesta. Los ítems de ejecución típica son muy diferentes a los de rendimiento máximo, así como también es diverso el proceso que lleva a un individuo a optar por alguna de las opciones del ítem. Mientras que en uno debe indicar qué hace usualmente (conducta típica), en el otro debe mostrar cuánto puede hacer (rendimiento máximo). Aunque a priori no existen razones para dejar de lado los modelos surgidos en el contexto de la medición educativa, también es factible suponer que adoptarlos de manera pragmática puede tener alguna consecuencia negativa aún desconocida (Chernyshenko, Stark, Drasgow & Roberts, 2007). Los factores que intervienen en la respuesta de un individuo a un ítem de ejecución típica son distintos a los que pueden afectar a los ítems de habilidades. Variables tales como la deseabilidad social, la ambigüedad-vaguedad de un enunciado o la complejidad sintáctica de un ítem adquieren mayor relevancia en la determinación de la calidad de un ítem de ejecución típica (Zickar & Ury, 2002).

El paulatino crecimiento de las aplicaciones de los modelos de la TRI a datos obtenidos con tests de ejecución típica mostró la necesidad de estudios minuciosos de las características del ítem y de la forma en que este es respondido (Ferrando & Demestre, 2008). Las investigaciones actuales están enfocadas a encontrar interpretaciones psicológicas específicas para los parámetros de los modelos de la TRI, especialmente en la medición de la personalidad (Zickar & Ury, 2002). Por esta razón, la postura pragmática fue perdiendo poco a poco su vigencia a medida que avanzaron los desarrollos en el área.

Un importante factor que afecta actualmente la expansión y puesta en práctica de la TRI en el contexto de los tests de ejecución típica es de orden operativo. Los procedimientos de estimación de los parámetros de los modelos requieren tamaños muestrales elevados para garantizar un bajo nivel de error en las estimaciones. Si bien los modelos más sencillos de la TRI pueden alcanzar estimaciones estables con 200 a 500 individuos, a medida que se incorporan más parámetros para describir el modelo, aumentan significativamente los requerimientos en cuanto a tamaño muestral. Diversas investigaciones han mostrado que para un óptimo funcionamiento de los procedimientos de estimación con ítems dicotómicos se requiere de al menos 500 aplicaciones. (Muñiz, 1997; Tsutakawa & Johnson, 1990). Para los modelos politómicos (más frecuentes en los tests de ejecución típica), este aspecto puede resultar más difícil, dado que se necesitará más cantidad de parámetros para describirlo (la cantidad de parámetros usados para describir los modelos de la TRI dependen de la cantidad de opciones de respuesta que tiene el ítem). En consecuencia,

los modelos para datos politómicos necesitarán tamaños muestrales considerablemente más grandes para alcanzar un criterio de convergencia aceptable en la estimación de los parámetros. Mediante un estudio de simulación, Reise y Yu (1990) encontraron que un tamaño de muestra mínimo de 500 puede ser aceptable para estimar los parámetros de ítems con cinco opciones de respuesta graduadas. Los estudios de simulación permiten establecer la potencia de los procedimientos estadísticos en determinadas condiciones, que pueden no ser iguales a las que se presentan en los estudios empíricos. Cuando se realizan investigaciones con datos reales pueden incidir otras variables ocultas que pueden obligar a trabajar con un tamaño muestral más elevado del que indica la simulación.

Reise (1999) resaltó como punto importante la poca accesibilidad a grandes muestras que se tiene en la investigación con tests de ejecución típica. Este aspecto se agudiza si se pretende abordar el campo de la psicología clínica. En contraposición, las muestras numerosas son más fácilmente asequibles y en un menor tiempo en contextos educativos, lo que garantiza una prolífica aplicación de los modelos de la TRI a tests de rendimiento máximo. Es conveniente señalar que Reise (1999) basa su justificación en la dificultad para conseguir muestras numerosas, pero no se refiere a la inversión económica que esto conlleva. Los investigadores de países menos desarrollados, como los de Latinoamérica, deben realizar un esfuerzo aún mayor por la falta de recursos financieros.

Consideraciones finales

La psicometría atraviesa un momento de enorme expansión teórica y aplicada sustentado en gran medida por los aportes de la TRI. Esta teoría ha permitido garantizar que la medición de rasgos psicológicos y las propiedades de los ítems se mantengan constantes más allá de los cambios en la composición del instrumento y del grupo normativo. Esto le confiere un estatus de mayor rigurosidad y objetividad a la medición de constructos psicológicos imposible de alcanzar desde la perspectiva clásica.

Pero los desarrollos de la TRI han tenido un crecimiento disparejo en función de los constructos evaluados. Mientras que la TRI ha conseguido cierto predominio en el análisis de pruebas de rendimiento, habilidades y aptitudes, los tests que indagan sobre actitudes, intereses o atributos de la personalidad continúan siendo explicados desde modelos como la TCT.

Es esencialmente el factor histórico el que determinó la marcada desproporción en las aplicaciones de la TRI en función del tipo de constructo. Su surgimiento y desarrollo en el contexto educativo obstaculizó su difusión como teoría psicométrica. La progresiva superación de estos factores implicó la difusión de la existencia de los modelos politómicos y de los beneficios que brinda la aplicación de la TRI al análisis de ítems de ejecución típica. Aunque queda un largo camino por recorrer, se ha avanzado bastante al respecto. En la actualidad, diversos libros sobre evaluación psicológica dedican uno o varios capítulos a las aplicaciones de los modelos de la TRI a variables de personalidad (Embretson & Hershberger, 1999; Embretson & Reise, 2000) y actitudes (Bond & Fox, 2001). Asimismo, los textos abocados a la medición de la personalidad han comenzado a dedicar un capítulo para comentar los avances producidos por la aplicación de la TRI (Morizot et al., 2007; Steinberg & Thissen, 1995).

A pesar de la preponderancia del factor histórico, no es despreciable el impacto que actualmente ocasionan los factores asociados a la aplicación, principalmente porque estos explican por qué la disparidad persiste hasta la actualidad y es altamente probable que permanezca también en las próximas décadas. El avance en la difusión de la TRI ha permitido el surgimiento de los primeros ensayos prácticos de los modelos a los datos conseguidos con tests de ejecución típica. Pero los intentos pragmáticos ya han

quedado atrás, la TRI ha demostrado su utilidad para modelizar tests de ejecución típica y en la actualidad se enfrenta a los desafíos que surgen de la especificidad de los constructos medidos.

Además de las ventajas técnicas que redundan en una medición de mejor calidad, la TRI también ofrece una valiosa comprensión del rasgo medido. Su aplicación devela la relación que existe entre los indicadores propuestos por los ítems y la capacidad de estos para discriminar en un valor específico del espectro del rasgo. Esta información le permite al investigador un estudio más profundo del atributo que mide y orienta respecto de una posible modificación en la teoría psicológica que fundamenta el constructo.

Los factores expuestos a la luz de la bibliografía psicométrica descubren que el uso de los modelos de la TRI para la medición con tests de ejecución típica siempre ha tenido obstáculos. Por otra parte, la contracara de los mismos factores muestra que las aplicaciones en tests de rendimiento máximo siempre estuvieron facilitadas tanto desde los aspectos teóricos como prácticos. Ambos hechos actúan de forma recíproca al momento de justificar la enorme desproporción en las aplicaciones según el tipo de constructo.

Muy probablemente la utilización de la TRI para la modelización de tests de ejecución típica nunca llegue a ser más frecuente que la de los tests de ejecución máxima. Los años de desarrollo que lleva de ventaja el campo educativo instan a que sus aplicaciones siempre estén un paso más adelante. Quizás resulte de utilidad dejar de depender de los avances generados en otros contextos y comenzar a fundar un campo de conocimiento que reconozca la especificidad de los constructos medidos con tests de ejecución típica; esto solo se podrá alcanzar si se establece un intercambio más fluido entre los especialistas en psicometría y los teóricos del constructo que se pretende medir.

Referencias

Aguado, D., Rubio, V. J., Hontangas, P. M. & Hernández, J. M. (2005). Propiedades psicométricas de un test adaptativo informatizado para la medición del ajuste emocional. *Psicothema, 17*, 484-491.

Aiken, L. R. (1999). *Personality assessment. Methods and practices.* Seattle: Hogrefe & Huber Publishers.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Baker. F. B. (2001). *The basics of item response theory.* Maryland: ERIC Clearinghouse on Assessment and Evaluation.

Bejar, I. (1977). An application of the continuous response level model to personality measurement. *Applied Psycholological Measurement, 1*, 509-521.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.

Cadavid, N., Delgado, A. R. & Prieto, G. (2007). Construcción de una escala de depresión con el modelo de Rasch. *Psicothema, 19*, 515-521.

Calero, M. D. & Padilla, J. L. (2007). Técnicas psicométricas: los tests. En R. Fernández-Ballesteros (dir.), *Evaluación psicológica. Conceptos, métodos y estudio de casos* (pp. 323-358). Madrid: Pirámide.

Carter, J. E. & Wilkinson, L. (1984). A latent trait analysis of the MMPI. *Multivariate Behavioral Research, 19,* 385-407.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F. & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

Chernyshenko, O. S., Stark, S., Drasgow, F. & Roberts, B. W. (2007). Constructing Personality Scales Under the Assumptions of an Ideal Point Response Process: Toward Increasing the Flexibility of Personality Measures. *Psychological Assessment, 19*, 88-106.

Childs, R. A., Dahlstrom, W. G., Kemp, S. M. & Panter, A. T. (2000). Item response theory in personality

assessment: A demonstration using the MMPI-2 Depression Scale. *Assessment, 7*, 37-54.

Cronbach, L. J. (1972). *Fundamentos de la exploración psicológica*. Madrid: Biblioteca nueva.

Damarin, F. (1970). A latent structure model for answering personal questions. *Psychological Bulletin, 73*, 23-40.

Embretson, S. E. & Hershberger, S. L. (1999). *The new rules of measurement*. Mahwah, NJ: Erlbaum.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Escorial, S. & Navas, M. J. (2007). Analysis of the gender variable in the Eysenck personality questionnaire revised scales using differential item functioning techniques. *Educational and Psychological Measurement, 67*, 990-1001.

Ferrando, P. J. (1994). Fitting response models to the EPI-A Impulsivity scale. *Educational and Psychological Measurement, 54*, 118-127.

Ferrando, P. J. & Chico, E. (2000). Adaptación y análisis psicométrico de la escala de deseabilidad social de Marlowe y Crowne. *Psicothema, 12*, 383-389.

Ferrando, P. J. & Demestre, J. (2008). Características de forma y contenido que predicen la capacidad discriminativa en ítems de personalidad: un análisis basado en la Teoría de Respuesta a los Ítems. *Psicothema, 20*, 851-856.

Flannery, W. P., Reise, S. P. & Widaman, K. F. (1995). An item response theory of the general and academic scales of the Self-Description Questionnaire II. *Journal of Research in Personality, 29*, 168-188.

Fraley, R. C., Waller, N. G. & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*, 350-365.

Gomez, R. & Fisher, J. W. (2005). Item response theory analysis of the spiritual well-being questionnaire. *Personality and Individual Differences, 38*, 1107-1121.

Gray-Little, B., Williams, V. S. L. & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, *23*, 443-451.

Hambleton, R. K. (1997). Perspectivas futuras y aplicaciones. En J. Muñiz (ed.), *Introducción a la Teoría de Respuesta a los Ítems* (pp. 203-213). Madrid: Pirámide.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Hendryx, M. S., Haviland, M. G., Gibbons, R. D. & Clark, D. C. (1992). An application of item response theory to alexithymia assessment among abstinent alcoholics. *Journal of Personality Assessment, 58*, 506-515.

Hernández, A., Espejo, B. & González-Romá, V. (2006). The functioning of central categories middle level and sometimes in graded response scales: Does the label matter? *Psicothema, 18*, 300-306.

Jones, L. V. & Thissen, D. (2007). A history and overview of psychometrics. En C. R. Rao & S. Sinharay (eds.), *Handbook of Statistics, 26: Psychometrics* (pp. 1-27). Amsterdam: North Holland.

Kline, P. (1983). *Personality: Measurement and theory*. London: Hutchinson.

Kuncel, R. B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Research, 34*, 545-563.

Lange, R. & Houran, J. (1999). Scaling MacDonald's AT-20 using item-response theory. *Personality and Individual Differences, 26*, 467-475.

López-Pina, J. A. (2005). Ítems politómicos vs. dicotómicos: Un estudio metodológico. *Anales de Psicología, 21*, 339-344.

Lord, F. M. (1952) *A theory of test scores. (Psychometric Monograph, 7)*. Iowa City, IA: Psychometric Society.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Martínez-Arias, M. R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Martínez-Arias, M. R. (1999). Los avances de la psicometría y la construcción de test. En F. Silva (ed.), *Avances en evaluación psicológica* (pp. 9-73). Valencia: Promolibro.

Morizot, J., Ainsworth, A. T. & Reise, S. P. (2007). Toward modern psychometrics. application of item

response theory models in personality research. En R. W. Robins, R. C. Fraley & R. F. Krueger (eds.), *Handbook of Research Methods in Personality Psychology* (pp. 407-423). New York: Guilford Press.

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

Nunnally, J. C. & Bernstein, I. J. (1995). *Teoría psicométrica* (3.ª ed.). México: McGraw-Hill.

Ostini, R. & Nering, M. L. (2005). *Polytomous item response theory models.* Newbury Park, CA: Sage.

Ozer, D. J. & Reise, S. P. (1994). Personality assessment. *Annual Review of Psychology, 45*, 357-388.

Prieto, G. & Delgado, A. R. (1999). Medición cognitiva de las aptitudes. En J. Olea, V. Ponsoda & G. Prieto (eds.), *Tests informatizados. Fundamentos y aplicaciones* (pp. 207-226). Madrid: Pirámide.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhague: The Danish Institute for Educational Research.

Reise, S. P. (1999). Personality measurement: issues viewed trough the eyes of IRT. En S. E. Embretson & S. L. Herschberger (eds.), *The new rules of measurement: What every psychology and educator should know* (pp. 219-241). Mahwah, NJ: Erlbaum.

Reise, S. P. & Henson, J. M. (2000). Computarizacion and adaptative administration of the NEO PI-R. *Assessment, 7* (4), 347-364.

Reise, S. P. & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93-103.

Reise, S. P. & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Reise, S. P. & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65,* 143-151.

Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Reiser, M. (1989). An application of the item response model to psychiatric epidemiology. En C. C. Clogg (ed.), *Sociological methodology* (pp. 271-307). Washington: American Sociological Association.

Rojas, A. J. & Pérez, C. (2001). *Nuevos modelos para la medición de actitudes*. Valencia: Promolibro.

Roskam, E. E. (1985). *Measurement and personality assessment*. Amsterdam: North Holland.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society. Recuperado el 4 de mayo del 2009, de http://www.psychometrika.org/journal/online/MN17.pdf

Schaeffer, N. C. (1988). An application of the item response theory to measurement of depression. En C. C. Clogg (ed.), *Sociological methodology* (pp. 271-307). Washington: American Sociological Association.

Singh, J., Howell, R. D. & Rhoads, G. K. (1990). Adaptive designs for Likert-Type data: An approach for implementing marketing surveys. *Journal of Marketing Research, 27*, 304-321.

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin, 28*, 754-763.

Steinberg, L. & Thissen, D. (1995). Item response theory in personality research. En P. E. Shrout & S. T. Fiske (eds.), *Personality research methods, and theory* (pp. 161-181). Hillsdale, NJ: Erlbaum.

Steinberg, L. & Thissen, D. (1996). Uses of item response theory and testlet concept in measurement of psychopatology. *Psychological Methods, 1*, 81-97.

Thissen, D., Steinberg, L., Pyszczynski, T. & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*, 211-226.

Tsutakawa, R. K., & Johnson, J. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371-390.

Waller, N. G., Tellegen, A., McDonald, R. P. & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary

validation of a negative emotionality scale. *Journal of Personality, 64*, 545-576.

Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton: Educational Testing Service.

Zickar, M. J. & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

Zickar, M. J. & Ury, K. L. (2002). Developing an interpretation of item parameters for personality items: content correlates of parameter estimates. *Educational and Psychological Measurement, 62*, 19-31.