SUMMA PSICOLÓGICA UST
2014, Vol. 11, Nº 2, 101-113

ISSN 0718-0446 (impresa)

Comparación del Modelo de Respuesta Graduada y la Teoría Clásica de Tests en una Escala de Confianza para la Matemática

A Comparison of the Graded Response Model and Classical Test Theory with a Scale of Confidence in Mathematics

Facundo Abal
Instituto de Investigaciones de la Facultad de Psicología, Universidad de Buenos Aires, Buenos Aires, Argentina
Consejo Nacional de investigaciones Científicas y Teológicas, Buenos Aires, Argentina
Sofía Auné
Instituto de Investigaciones de la Facultad de Psicología, Universidad de Buenos Aires, Buenos Aires, Argentina
Agencia Nacional de Promoción Científica y Tecnológicas, Buenos Aires, Argentina
Horacio Attorresi
Instituto de Investigaciones de la Facultad de Psicología, Universidad de Buenos Aires, Buenos Aires, Argentina

(Rec: abril 2014 – Acep: noviembre 2014)

Resumen

Se aplicó el Modelo de Respuesta Graduada (MRG) de la Teoría de Respuesta al ítem (TRI) y la Teoría Clásica de Test (TCT) al análisis de ítems de una escala de Confianza para la Matemática (Abal, 2013). La prueba mide la capacidad percibida por un estudiante universitario para operar eficazmente con símbolos y fórmulas, aprender y aprobar la asignatura matemática u otras afines. La prueba consta de 8 ítems con formato de respuesta Likert de 6 opciones. Participaron 1875 estudiantes de Psicología de la Universidad de Buenos Aires, Argentina. Se verificó la condición de unidimensionalidad requerida por el MRG. El ajuste del MRG fue satisfactorio para todos los ítems. El análisis clásico incluyó el estudio de las frecuencias de respuesta, estadísticos descriptivos del ítem y correlación ítem-test corregida. El coeficiente de confiabilidad marginal de la TRI fue de .91 y el alfa de Cronbach fue .90. Se encontraron correlaciones elevadas entre: a) la media del ítem y los parámetros de localización centrales del MRG, b) la correlación ítem-test corregida y los parámetros de discriminación y c) entre los escalamientos de los individuos realizados desde la TRI y la TCT. Estos resultados aportan evidencias de validez basadas en la estructura interna del instrumento.
Palabras clave: Confianza para la Matemática, Teoría de Respuesta al Ítem, Modelo de Respuesta Graduada, Teoría Clásica de Test.

Abstract

The Graded Response Model (GRM) of Item Response Theory (IRT) and Classical Test Theory (CTT) were applied to the analysis of items from a scale of Confidence in Mathematics (Abal, 2013). This scale measures the ability perceived by university student to operate effectively with symbols and formulas, to solve problematic situations, to learn and pass mathematics or related subjects. The scale comprises 8 items in polytomous response format (6-point Likert-type). The sample was made up by 1875 students of the Psychology school of Buenos Aires University, Argentina. The unidimensionality assumption required by the GRM was confirmed. The GRM fitted to data satisfactorily for all items. Location and discrimination parameters showed predictable values. Classical item analysis involved the examination of response frequencies, item descriptive statistics and corrected item-test correlations. The marginal reliability coefficient obtained from IRT was .91 and Cronbach's alpha was .90. High correlations were found between: a) item means and central location parameters of GRM, b) corrected item-test correlations and discrimination parameters, and c) IRT and CTT individual scores. The finding provides validity evidences based on the internal structure of scale.
Key words: Confidence in Mathematics, Item Response Theory, Graded Response Model, Classical Test Theory.

Correspondencia dirigida a: Facundo Juan Pablo Abal. Universidad de Buenos Aires, Viamonte 430/44 C1053ABJ, Ciudad de Buenos Aires, Argentina. Tel: (054 11) 4510 1100. E-mail: fabal@psi.uba.ar

Introducción

Desde hace ya varios años la Psicometría mundial se encuentra atravesando un período de transición tanto en lo que se refiere a sus aspectos teóricos como aplicados (Ayala, 2009; De Mars, 2010; Hambleton, 2004). La bibliografía sobre Teoría de los Tests de los últimos cuarenta años ha registrado un desplazamiento gradual hacia teorías y técnicas de medición psicológica superadoras de la Teoría Clásica de Tests (TCT) (Jones & Thissen, 2007; Muñiz, 1997). Una gran parte de estas nuevas contribuciones se asienta en las innovaciones teóricas y tecnológicas introducidas por la Teoría de Respuesta al Ítem (TRI).

La TCT surgió del modelo lineal de puntuaciones establecido por Spearman (1904) y alcanzó su formalización más precisa en la obra de Novick (1966). Su formulación matemática es bastante simple y supone que el puntaje observado de una persona en un test es el resultado de la suma del valor real (puntaje verdadero) y el error de medición. A pesar de su extendida aplicación, nunca faltaron las críticas al modelo y a los supuestos que lo fundamentan. Los principales problemas conceptuales ya habían sido aludidos muy tempranamente por Thurstone (1928): a) las propiedades de los ítems y del test dependen de los individuos utilizados para establecerlas en la muestra de estandarización y b) la medición de las variables depende del instrumento. Son estas limitaciones, entre otras, las que propiciaron el desarrollo de otras perspectivas superadoras como la TRI.

Las ideas fundamentales sobre la TRI ya se habían desarrollado a nivel teórico a mediados de siglo XX, pero la complejidad de los cálculos hacía extremadamente dificultosa su puesta en práctica. Sólo con el avance de la informática se vislumbraron las posibilidades de su aplicación (Lord, 1980). La TRI permitió llevar a la práctica una prolífera producción de modelos teóricos que revolucionaron la construcción y administración de los test utilizados en el campo de la evaluación psicológica y educativa (Abad, Olea, Ponsoda & García, 2011; Muñiz, 1997).

El objetivo sustancial de la TRI es la elaboración de tests con propiedades invariantes entre poblaciones. Wright (1968) empleó la denominación tests libres de muestra (*sample-free test*) para resumir este concepto. Esto implica que la TRI permite obtener estimaciones no sesgadas de las propiedades de los reactivos aún partiendo de muestras no representativas de la población (Embretson & Reise, 2000). Otra ventaja de la TRI por sobre la TCT es que permite efectuar mediciones invariantes más allá de la composición del instrumento. El nivel de rasgo latente que presenta un individuo es producto de una estimación obtenida a partir del patrón de respuestas manifestado en un conjunto de ítems. Si se varía el conjunto de ítems utilizado también se mantiene la puntuación estimada aunque eventualmente hayan cambiado las propiedades psicométricas de los reactivos. Finalmente, Hambleton y Swaminathan (1985) también rescataron la importancia de las medidas locales de precisión que proporciona la TRI. En el marco de esta teoría, es posible obtener información acerca de la precisión con que el test y cada ítem en particular estiman cada nivel del rasgo latente del individuo. Esto remite, por tanto, al concepto de fiabilidad de perspectiva clásica; sólo que en la TRI un ítem o un test no será más o menos confiable en términos absolutos, sino para determinados niveles de la escala.

Ahora bien, aunque la TRI presenta grandes ventajas, la TCT no ha perdido su vigencia. La razón por la cual sigue siendo ampliamente utilizada como soporte teórico para la construcción de tests es que carece de supuestos exigentes. La aplicación del modelo lineal se torna accesible a la mayoría de los datos empíricos. Como afirmó Muñiz (1994), la baja exigencia es a la vez la fuerza y la debilidad que tiene la TCT en tanto que sacrifica su capacidad predictiva arrojando conclusiones extremadamente generales. Sin embargo, el estudio de la calidad de los ítems en el marco de la TCT presenta las limitaciones que surgen del uso de indicadores globales en la evaluación de la calidad de los ítems. Este análisis resulta de gran utilidad pero está lejos de ser exhaustivo por lo que siempre es fructífero profundizar en el estudio del funcionamiento de los reactivos.

La coexistencia de la TCT y la TRI no implica su antagonismo. Lejos de competir, los modelos psicométricos se superponen teóricamente y pueden utilizarse de forma complementaria para realizar un análisis más exhaustivo de la calidad o del funcionamiento del test (Hulin, Drasgow & Parsons, 1983). Como aseguró Lord (1980), la TRI no contradice ni desestima los análisis efectuados desde la TCT. Incluso, se han reportado numerosos hallazgos que muestran la correspondencia entre indicadores psicométricos de la TCT y sus equivalentes conceptuales en el marco de la TRI (Barbero, Prieto, Suárez & San Luis, 2001; Kramp, 2008).

A pesar del fuerte desarrollo que ha tenido la TRI en los último años, su aplicación a tests de comportamiento típico resulta poco frecuente (Abal, Lozzia, Aguerri, Galibert & Attorresi, 2010; Morizot, Ainsworth & Reise, 2007). Por esta razón, el objetivo general de este

trabajo es analizar los ítems de una prueba que mide el constructo Confianza para la Matemática aplicando un modelo politómico de la TRI. Asimismo, se busca establecer una comparación de los resultados obtenidos con los indicadores de calidad psicométrica empleados tradicionalmente desde la perspectiva clásica.

La confianza para la Matemática forma parte de la dimensión afectiva de la enseñanza y aprendizaje de esta materia. Esta dimensión incluye un extenso rango de creencias, actitudes y emociones que intervienen en los procesos de adquisición del conocimiento matemático, pero que son considerados como una entidad diferente a la pura cognición (Gómez-Chacón, 2005; Martínez-Padrón, 2008). La *confianza* se caracteriza por un conjunto de percepciones y creencias del estudiante sobre sus posibilidades y dificultades para responder a las habilidades requeridas en la actividad matemática (Abal, 2013). Se trata de un constructo que ha recibido múltiples denominaciones y que tradicionalmente ha sido estudiado en el marco de la actitud hacia la Matemática (Hernández, 2011; Tapia & Marsh, 2004).

El instrumento para la medición de la confianza se encuentra en proceso de elaboración. Se construyó considerando una población específica de estudiantes de carreras humanístico-sociales, con el fin de describir las características particulares de este grupo y, a futuro, pensar en estrategias didácticas para materias de corte cuantitativo de estas carreras. Se ha llevado adelante una depuración primaria y secundaria exigente en etapas previas de la investigación. Los ítems fueron sometidos a la crítica de jueces expertos y a una serie de estudios piloto donde se recogieron evidencias de validez de contenido y aparente (Abal, 2013).

Considerando el formato de respuesta tipo Likert de la Escala de Confianza para la Matemática se decidió aplicar el Modelo de Respuesta Graduada (MRG) de Samejima (1969, 1997). Esta elección está en línea con hallazgos de otros autores, los cuales encontraron que la aplicación del MRG mostraba ventajas por sobre otros modelos politómicos y dicotómicos de la TRI (Asún & Zúñiga, 2008; Baker, Rounds & Zevon, 2000; King, King, Fairbank, Schlenger, & Surface, 1993).

Samejima (1997) propuso el MRG como una extensión del Modelo Logístico de dos parámetros de Birnbaum (1968). Para describir el funcionamiento de un ítem, el MRG utiliza un parámetro de discriminación a_i y una serie de parámetros de umbral b_{ik} ($k = 1,\ldots, m$) que se ubican entre las categorías k contiguas del ítem politómico ($k = 0, \ldots, m$). El parámetro b_{ik} refleja, con la misma métrica del rasgo latente θ, cuánto rasgo es necesario para tener una probabilidad igual a .50 de responder la opción k o una superior. La autora definió las *Curvas Características Operantes* $P^*_{ik}(\theta)$ como una función que representa la probabilidad de que la respuesta de un examinado al ítem i esté en o por encima de un umbral b_{ik} en función del nivel del rasgo latente θ a partir de la ecuación:

$$P^*_{ik}(\theta) = \frac{1}{1 + e^{-1.7 a_i (\theta - b_k)}};\ k = 1, \ldots, m$$

Las Curvas Características Operantes constituyen sólo un paso intermedio, dado que el interés está centrado en obtener las *Curvas Características de las Categorías de Respuesta del Ítem* (CCCRI); (Samejima, 1969). Estas últimas representan la probabilidad que tiene un individuo de optar por la categoría k en función de su nivel en el rasgo latente θ y se denotan como $P_{ik}(\theta)$. Dada la categoría k del ítem i y la categoría inmediata superior $(k+1)$ es posible calcular $P_{ik}(\theta)$ como una sustracción de las probabilidades acumuladas a derecha:

$$P_{ik}(\theta) = P^*_{ik}(\theta) - P^*_{i;k+1}(\theta)$$

Por definición, la probabilidad acumulada de responder en o por encima de la categoría más baja es $P^*_{i0} = 1$ mientras que la probabilidad de responder por encima de la última categoría es $P^*_{i(m+1)} = 0$. Por ende,

$$P_{i0}(\theta) = P^*_{i0}(\theta) - P^*_{i1}(\theta) = 1 - P^*_{i1}(\theta)$$

$$P_{ih}(\theta) = P^*_{ih}(\theta) - P^*_{i;h+1}(\theta) \text{ para } h \text{ tal que } 1 \le h \le m\text{ -}1$$

$$P_{im}(\theta) = P^*_{im}(\theta) - P^*_{i;m+1} = P^*_{im}(\theta) - 0 = P^*_{im}(\theta)$$

A partir del MRG también es posible obtener medidas locales de precisión, las cuales se hacen operativas mediante la Función de Información (FI) del test y de los ítems. La FI de un ítem muestra la precisión con que el reactivo mide el rasgo latente a lo largo de todo su rango de valores. Esto permite identificar para qué niveles de rasgo el ítem resulta más o menos confiable. En virtud de la aditividad de las FIs de los ítems para todos los niveles del rasgo, es posible obtener una FI del test completo así como una función de error típico de la estimación. La FI de un test $I(\theta)$ con j ítems y su error típico de estimación $Se(\theta)$ se calculan como (Muraki & Bock, 2003):

$$\therefore I(\theta) = \sum_{i=1}^{j} \sum_{k=0}^{m} \frac{1.7^2 a^2 \left\{ P^*_{ik}(\theta)\left[1 - P^*_{ik}(\theta)\right] - P^*_{i;k+1}(\theta)\left[1 - P^*_{i;k+1}(\theta)\right] \right\}^2}{P_{ik}(\theta)}$$

$$Se\ (\theta) = \frac{1}{\sqrt{I(\theta)}}$$

En el marco de la TRI también es posible realizar una estimación de la precisión global con que mide un instrumento a través del coeficiente de confiabilidad marginal (Thissen, 2003). Este indicador se calcula como un promedio de la imagen de la FI del test a través de todos los niveles del rasgo ponderado por la proporción de individuos que corresponden a cada nivel. Su efectividad para caracterizar la precisión de la prueba está sujeta a una FI del test relativamente uniforme. Bajo estas circunstancias, su valor se aproxima y admite una interpretación idéntica al alfa de Cronbach obtenido desde la TCT.

Método

Participantes

Se contó con la colaboración de 1875 estudiantes de ambos sexos (82% fueron mujeres) de segundo año de la Facultad de Psicología de la Universidad de Buenos Aires. Se utilizó un muestreo no probabilístico por conveniencia. La edad de los participantes osciló entre 18 y 62 años, con una media de 22.7 años (*DE* = 6.33), una mediana de 20 y una amplitud semi-intercuartil de 2 años. La distribución de la edad presentó una marcada asimetría positiva con un índice de 3.06. Para favorecer la sinceridad en las respuestas y dar garantías de confidencialidad los individuos contestaron de manera anónima.

Instrumento

Escala de Confianza para la Matemática (Abal, 2013). Consta de ocho enunciados con formato de respuesta politómica de seis opciones tipo Likert (*Totalmente en desacuerdo, En desacuerdo, Más bien en desacuerdo, Más bien de acuerdo, De acuerdo y Totalmente de acuerdo*). Los reactivos pueden verse en la figura 1. Los estudiantes con altos niveles de confianza se perciben seguros y eficaces ante situaciones donde deben aplicar nociones de matemática. Creen poseer conocimientos sólidos, ser capaces de incorporar con rapidez nuevos conceptos y presentan niveles de ansiedad relativamente bajos que no dificultan la realización de las actividades matemáticas. En cambio, los estudiantes con bajos niveles en esta variable se reconocen con dificultades para incorporar, recuperar y aplicar lo aprendido en Matemática. Dudan de su habilidad para resolver problemas y experimentan niveles de tensión que dificultan su capacidad para pensar con claridad.

Procedimiento

Se adoptó un diseño acorde con los requisitos de un estudio instrumental de corte psicométrico (Montero & León, 2007). Los estudiantes respondieron el instrumento de forma autoadministrada y en grupos de aproximadamente 30 personas coordinados por uno de los miembros del equipo de investigación. La administración se efectuó durante el horario de clase y la participación fue voluntaria. Previa aplicación de la prueba se brindó una exposición motivadora que resumió la finalidad de la actividad y la futura utilización de los datos. También se aclaró que la colaboración no tenía consecuencias ni negativas ni positivas con respecto a su rendimiento académico. Si bien los individuos contestaron sin tiempo límite, ninguno de los grupos tardó más de 10 minutos.

Análisis de datos

Se analizó la dimensionalidad del constructo mediante un estudio exploratorio de la estructura factorial de los datos mediante el programa Factor 9.2 (Lorenzo-Seva & Ferrando, 2006). Dado el carácter ordinal de la escala Likert usada para los ítems, se aplicó como procedimiento para la extracción de factores el método de mínimos cuadrados simple sobre la matriz de correlaciones de policóricas.

La aplicación de la TRI se efectuó operando el programa MULTILOG (Thissen, 2003). Para la estimación de los parámetros de los ítems y de las personas se aplicó el método de Máxima Verosimilitud. Este programa también permitió obtener medidas de precisión basadas en la TRI: la Función de Información del test y el coeficiente de confiabilidad marginal. Como indicador del ajuste de los datos del ítem al modelo, el programa brinda las proporciones observadas y esperadas de elección para cada categoría de respuesta. Dado que esta información como única evidencia es limitada (Revuelta, Abad & Ponsoda, 2006), se adicionaron indicadores indirectos recomendados por Rubio, Aguado, Hontangas y Hernández (2007): a) la convergencia en la estimación de los parámetros utilizando una cantidad razonable de iteraciones, b) la estimación de parámetros

con valores acordes a lo esperable, c) la existencia de errores estándares relativamente bajos, d) evidencia empírica de la invarianza de los parámetros.

En el marco de la perspectiva clásica se obtuvieron las distribución de frecuencias porcentuales de las categorías de respuesta politómica, estadísticos descriptivos (media, desvío típico, asimetría y curtosis) e índices de discriminación (correlación ítem-test corregida). La confiabilidad se analizó del instrumento a partir del alfa de Cronbach. No obstante, dado los cuestionamientos que ha recibido este coeficiente en los últimos años (Elosua & Zumbo, 2008; Sijtsma, 2009), se ha complementado este indicador con el *greatest lower bound* (glb) (Ten Berge, Snijders & Zegers, 1981).

Para estudiar las relaciones entre los estadísticos de la TCT y los parámetros de la TRI se estudió la correlación de: a) los parámetros de localización de los ítems y el puntaje promedio de cada uno de los ítems en la escala Likert, b) el parámetro *a* con la correlación item-test corregida. Finalmente, para estudiar las diferencias en el escalamiento realizados desde la TCT y la TRI, se estudió la correlación entre los parámetros θ estimados para cada individuos y los puntajes brutos calculados en la TCT como suma simple de las puntuaciones dadas a los ítems.

Resultados

Dimensionalidad

Para analizar la dimensionalidad del constructo se obtuvieron previamente indicadores que corroboraron la factibilidad de un estudio factorial. La medida de la adecuación muestral de Kaiser-Meyer-Olkin (*KMO*) presentó un valor de .93, lo cual muestra una interrelación satisfactoria entre los ítems y, además, una excelente adecuación muestral de los datos. Mediante la prueba de esfericidad de Bartlett se pudo rechazar la hipótesis nula de esfericidad ($\chi^2 = 7187.1$; $gl = 28$; $p<.0001$). En consecuencia, se rechazó la incorrelación lineal entre las variables, indicando que resulta apropiado aplicar el estudio factorial a este conjunto de datos.

Si bien se partió de un criterio teórico que a priori supuso la unidimensionalidad del constructo analizado (Hair, Anderson, Tatham & Black, 1999), se consideró la regla Kaiser (1960) y el análisis paralelo propuesto por Timmerman y Lorenzo-Seva (2011) para justificar la extracción de un único factor. El porcentaje de varianza total asociado al primer autovalor de 4.9 fue del 61.6% mientras que el porcentaje asociado al segundo autovalor de 0.6 fue del 7.4%. Las saturaciones factoriales de los ítems en el primer factor fueron adecuados y oscilaron entre .69 y .81. Estos resultados cumplen con los criterios propuestos por Kaiser (1960), Carmines y Zeller (1979) y Martínez-Arias (1995) para tener una aproximación aceptable al supuesto de unidimensionalidad requerido para la aplicación del MRG.

Aplicación del MRG

El proceso iterativo de estimación de los parámetros de los ítems alcanzó el criterio de convergencia de .001 en el ciclo 24. Para cada ítem se estimaron un parámetro discriminación (*a*) y cinco parámetros de localización (b_1, b_2, b_3, b_4 y b_5). Como los reactivos de la prueba fueron respondidos en una escala Likert de seis categorías y estaban redactados de forma inversa, el parámetro b_1 debe ser entendido como un valor de umbral que separa las categorías *Totalmente de acuerdo* y *De acuerdo*. Por lo tanto, b_1 se trata del mínimo valor de rasgo necesario para tener una probabilidad mayor a .50 de contestar la opción *De acuerdo* o una superior. Así también, el parámetro b_2 separa las categorías *De acuerdo* de *Más bien de acuerdo* y de igual modo ocurre para los sucesivos parámetros y categorías. En la tabla 1 se resumen los resultados del proceso de calibración en la que se incluyen los valores de los parámetros (*a* y b_k) de cada ítem, sus errores de estimación y estadísticos descriptivos. La figura 1 exhibe una matriz con las CCCCR de los ocho ítems de la prueba.

El estudio del ajuste del modelo a los ítems a partir de los residuos de las proporciones observadas y esperadas según el MRG reveló que las discrepancias entre unas y otras no fueron significativas. En consecuencia, los datos que proporcionó MULTILOG permitieron corroborar un ajuste adecuado de todos los reactivos que componen la prueba. En la misma línea, los indicadores indirectos también mostraron evidencias como para suponer el ajuste del MRG a los datos. Los 48 parámetros de los ítems fueron estimados en una cantidad de ciclos adecuada y presentaron un rango razonable de valores. El máximo error de estimación (*Se*) registrado fue de 0.10 y perteneció al último parámetro de umbral del ítem 7. Se puede concluir entonces que estos errores resultaron bajos, lo que aporta más evidencias para confiar en la estimación de los parámetros alcanzada. En líneas generales, los errores de estimación de los umbrales extremos b_1 y b_5 tendieron a ser más elevados que los centrales. La explicación de este resultado se puede hallar en las distribuciones de frecuencias de los

Tabla 1.

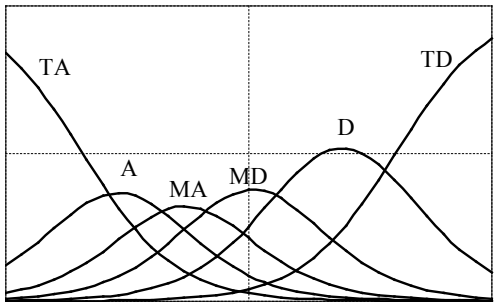
Parámetros de discriminación y de localización con sus errores de estimación.

Ítem	a (Se)	b_1 (Se)	b_2 (Se)	b_3 (Se)	b_4 (Se)	b_5 (Se)	b_{prom}
1	1.75 (0.07)	-2.05 (0.09)	-1.17 (0.06)	-0.41 (0.04)	0.50 (0.05)	1.81 (0.08)	-0.26
2	2.06 (0.08)	-1.71 (0.07)	-0.97 (0.05)	-0.21 (0.03)	0.46 (0.04)	1.60 (0.07)	-0.17
3	2.05 (0.08)	-1.38 (0.06)	-0.43 (0.03)	0.36 (0.04)	1.15 (0.05)	2.14 (0.09)	0.37
4	1.91 (0.08)	-1.72 (0.07)	-0.99 (0.05)	-0.23 (0.03)	0.57 (0.05)	1.74 (0.08)	-0.13
5	2.45 (0.08)	-1.36 (0.05)	-0.70 (0.04)	-0.08 (0.04)	0.70 (0.04)	1.68 (0.06)	0.05
6	2.04 (0.07)	-2.04 (0.08)	-1.20 (0.05)	-0.36 (0.03)	0.66 (0.04)	1.98 (0.09)	-0.19
7	2.21 (0.08)	-1.66 (0.06)	-0.67 (0.04)	0.20 (0.04)	1.08 (0.05)	2.29 (0.10)	0.25
8	2.65 (0.09)	-1.61 (0.06)	-0.92 (0.04)	-0.14 (0.03)	0.61 (0.04)	1.66 (0.06)	-0.08
Media	2.14	-1.69	-0.88	-0.11	0.72	1.86	
DE	0.27	0.24	0.25	0.25	0.24	0.23	
Mínimo	1.75	-2.05	-1.20	-0.41	0.46	1.60	
Máximo	2.65	-1.36	-0.43	0.36	1.15	2.29	

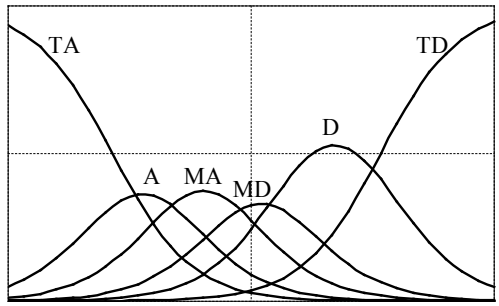
ítems (tabla 2). Las categorías que denotaban niveles de acuerdo intermedios con los enunciados fueron más elegidas por los individuos que las categorías extremas. Esto redundó en una mayor precisión en la estimación de los parámetros centrales (b_2, b_3 y b_4).

Se obtuvieron evidencias empíricas de la invarianza de las mediciones del rasgo latente respecto del test. Para esto se dividió la escala de forma aleatoria en dos subconjuntos de cuatro reactivos. La correlación de los θ estimados para cada grupo de ítems resultó de .83. Este índice puede considerarse elevado contemplando que los θ fueron estimados con un conjunto muy reducido de reactivos y esto propicia un aumento en el error de estimación. También se encontraron evidencias de la invarianza de las propiedades de los ítems respecto del grupo normativo. Las correlaciones de los parámetros de los ítems estimados a partir de dos subconjuntos de individuos de la muestra (N_1 = 938 y N_2 = 937) asignados de manera aleatoria adoptaron índices de .92 (parámetro b_5), .94 (parámetro a), .95 (parámetro b_1) y .98 (parámetros b_2, b_3 y b_4).

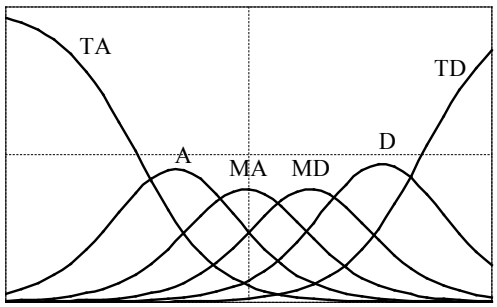
Los parámetros de discriminación revelaron que la escala presenta una capacidad discriminadora elevada con un valor promedio de a de 2.14 (DE = 0.27). El



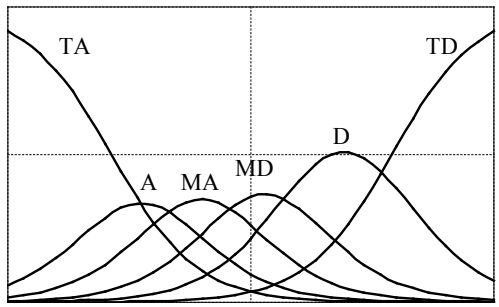
1. Me cuesta incorporar los conceptos matemáticos, aunque dedique mucho tiempo a estudiarlos.



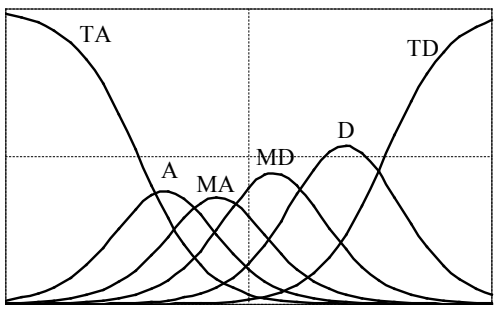
2.- Cada vez que enfrento un problema que se resuelve matemáticamente me pongo tenso.



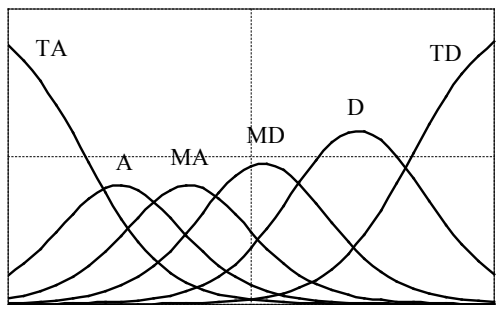
3. No soy muy rápido/a para hacer razonamientos matemáticos.



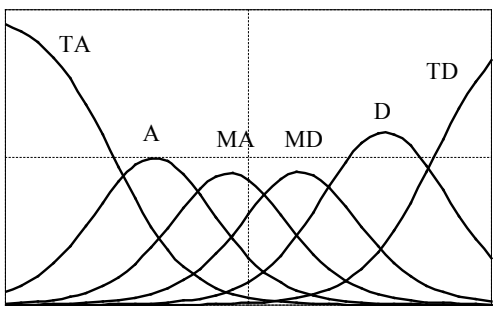
4. Cuando resuelvo problemas usando números siempre pienso que me voy a equivocar.



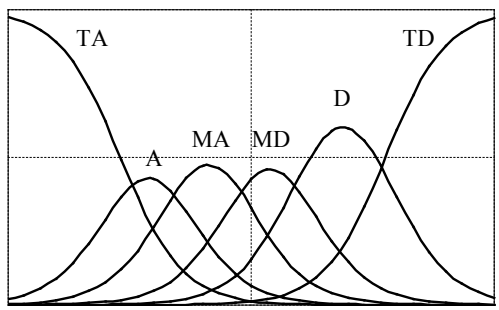
5. No me siento seguro/a cuando me presento a rendir exámenes donde tendré que hacer cálculos.



6. Cuando veo números en un planteo me cuesta pensar con claridad.



7. En general, me resulta difícil comprender las deducciones de los teoremas matemáticos.



8. Sé que tendré dificultades para aprobar las materias que se basen en los conocimientos de Matemática.

Figura 1.

Curvas Características de las Categorías de Respuesta de los ítems. TD = Totalmente en desacuerdo; D = En desacuerdo; MBD = Más bien en desacuerdo; MBA = Más bien de acuerdo; A = De acuerdo; TA = Totalmente de acuerdo.

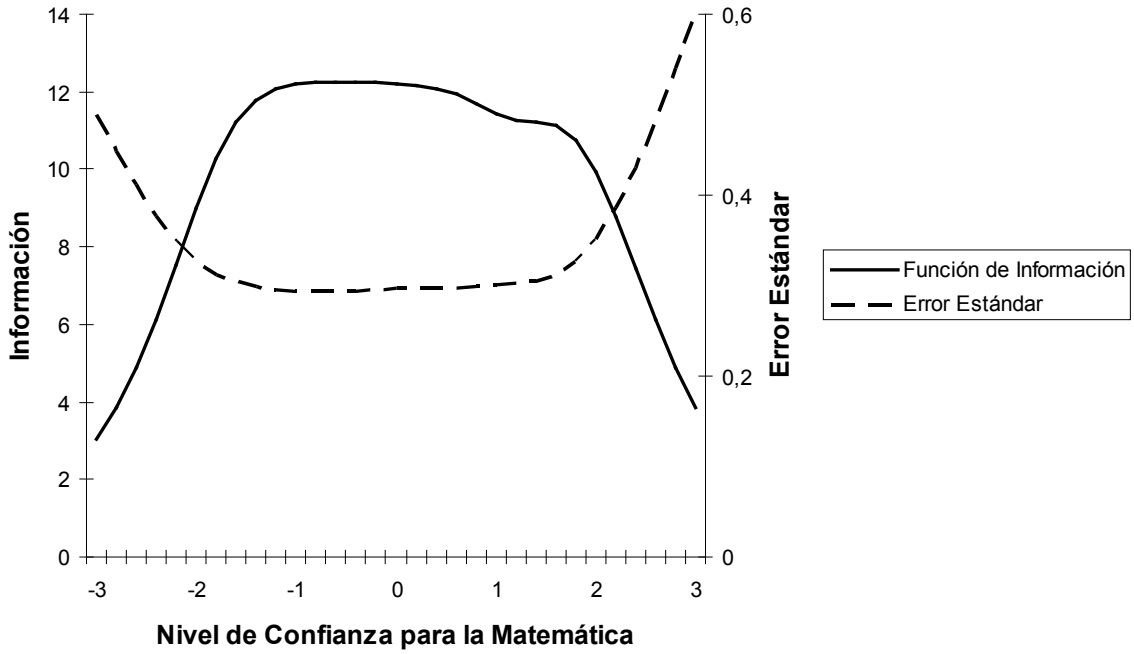


Figura 2.
Función de Información del Test y Error Estándar.

parámetro más alto (a = 2.65) perteneció al ítem 8 mientras que el más bajo (a =1.75) fue del ítem 1. Aún siendo este último el parámetro a más pequeño, resultó elevado con respecto a la clasificación de discriminación de Baker (2001). Los parámetros de umbral oscilaron entre -2.05 (b_1 ítem 1) y 2.29 (b_5 ítem 18), lo que permite abarcar la medición precisa del un rasgo latente en un rango amplio. Las distancias entre los valores de los b_k fueron considerables en los ocho reactivos. Todas las categorías resultaron máximamente probables en alguna parte recorrido del rasgo. Esto permite demostrar la eficacia de la escala Likert para discriminar distintos rangos de la variable, por lo que no parece justificada la reducción de la cantidad de categorías de respuesta. Los θ estimados tuvieron un promedio de 0.001 (*DE* = 0.95) siendo -2.53 el mínimo y 2.69 el máximo. La distribución resultó simétrica (*As* = 0.02), ligeramente platicúrtica (*Cr* =-0.19) y se ajustó al modelo normal (Prueba de Kolmogorov-Smirnov, Z = 0.02; p =.190).

La última columna de la tabla 1 muestra la posición relativa de cada ítem en la escala del rasgo latente al considerar al reactivo globalmente (promedio de los parámetros b_k del ítem). Cinco de los ocho b_{prom} quedaron ubicados por debajo de la media. Sin embargo, aunque existe una tendencia de los umbrales a posicionarse en los niveles medio-bajos del rasgo latente, esto resulta balanceado por los valores positivos de los b_4 y b_5 de todos los reactivos. Al indicar una localización global del ítem en la escala θ, los b_{prom} permiten ordenar los contenidos reflejados por los indicadores de los reactivos en función del nivel de Confianza que demandan. Por ejemplo, comparando los valores de b_{prom} más extremos se podría afirmar que un estudiante necesitará un mayor nivel de θ para tender a estar en desacuerdo con el ítem 3 (b_{prom}=0.37) que con el ítem 1 (b_{prom}= -0.26). Esto mismo tiene su correlato en el análisis de los indicadores. Un estudiante con un nivel de confianza medio-bajo para la Matemática podría considerar que, aunque le cueste, sería capaz de incorporar los conceptos si dedica tiempo de estudio (ítem 1). No obstante, necesitará de una mayor confianza para reconocer que es rápido para hacer razonamientos matemáticos (ítem 3).

Más allá de las diferentes localizaciones de los ítems consideradas por sus b_{Prom}, es posible apreciar en la figura 1 que la mayoría de los reactivos presentan CCCCR con formas parecidas. Esto es, se destaca la curva de la opción *En desacuerdo* por alcanzar una

Tabla 2.
Índices clásicos.

Ít.	Índice de atracción						Estadísticos descriptivos			*r*	Alfa si se elimina el ítem
	TD (%)	D (%)	MBD (%)	MBA (%)	A (%)	TA (%)	Media (DE)	As	Cr	i-T	
1	10	27	26	18	12	7	3.82 (1.40)	-0.37	-0.71	.63	.891
2	11	26	20	20	13	10	3.74 (1.49)	-0.27	-0.92	.69	.885
3	5	14	20	24	22	15	3.13 (1.42)	0.21	-0.86	.68	.886
4	10	25	23	20	12	10	3.7 (1.47)	-0.28	-0.86	.66	.889
5	9	20	24	19	15	13	3.48 (1.51)	-0.13	-0.99	.72	.883
6	7	25	30	21	11	6	3.77 (1.29)	-0.35	-0.50	.66	.888
7	4	16	24	26	20	10	3.29 (1.32)	0.04	-0.78	.69	.886
8	9	23	24	23	13	9	3.64 (1.41)	-0.20	-0.79	.74	.881

Nota. Ít = Número de ítem; TD = Totalmente en desacuerdo; D = En desacuerdo; MBD = Más bien en desacuerdo; MBA = Más bien de acuerdo; A = De acuerdo; TA = Totalmente de acuerdo; As = Asimetría; Cr = Curtosis; *r* i-T = Correlación ítem-test corregida.

mayor probabilidad de elección que el resto de las categorías centrales. La probabilidad que tiene una categoría *k* de ser elegida depende de la proximidad o lejanía de los parámetros b_{k-1} y b_k. En todos los ítems se corrobora que la distancia entre los umbrales b_4 y b_5 es mayor a las registradas entre los otros umbrales sucesivos. En cambio, las distancias entre los otros parámetros *b* fueron relativamente homogéneas y, por ende, las curvas mostraron una probabilidad de elección similar. El ítem 3 también responde a este patrón aunque de manera más sutil. En este caso, los parámetros *b* involucrados para *En desacuerdo* (b_5–b_4 = 0.99) estuvieron prácticamente igual de espaciados que los de la categoría *De acuerdo* (b_2–b_1 = 0.95).

La figura 2 muestra la Función de Información del Test (FIT) y su correspondiente Error Estándar. La FIT alcanzó su máxima información de 12.3 en un nivel de Confianza para la Matemática de -0.4. Para el mismo θ se registró el mínimo error estándar de 0.29. La FIT se mantiene relativamente elevada y constante en un amplio rango del rasgo latente (entre -1.6 y 2). Este resultado reveló que la prueba funciona adecuadamente para un rango de valores del rasgo bastante amplio, tanto por encima como por debajo de la media. Alcanzar una información pareja a lo largo de gran parte del espectro del rasgo latente garantiza que las estimaciones de los niveles de θ se lleven a cabo con un error estándar similar. Por ende, la medición es más justa para mayor cantidad de individuos. La ligera depresión de la FIT que se observa entre 0.8 y 1.6 es una consecuencia de la escasa cantidad de parámetros de umbral que permitan discriminar en esa región. En la tabla 1 se puede apreciar que los b_4 de los ítems 3 y 7 se encuentran en este rango de la variable y no es hasta un θ = 1.60 (b_5 del ítem 2) que aparece el siguiente parámetro de localización en la escala del rasgo.

Dado que la FIT se mostró relativamente uniforme a lo largo del recorrido del rasgo latente, resulta adecuado el análisis de la confiabilidad marginal de la prueba. Este coeficiente fue elevado alcanzando un valor de .91. Esto permite concluir que la confiabilidad del instrumento es satisfactoria.

Comparación de la TRI con los estadísticos de la TCT

En la tabla 2 aparecen las distribuciones de frecuencias de las categorías de respuesta y estadísticos descriptivos de los ítems. Los índices de asimetría y curtosis adoptaron valores absolutos inferiores a 1. Los índices de discriminación obtuvieron valores adecuados entre .63 y .74.

Las correlaciones entre los parámetros de los ítems de la TRI y los indicadores de la TCT respondieron a lo esperable. El parámetro *a* del MRG se encontró fuertemente asociado con la correlación item-test corregida (r = .97; p<.001). Los parámetros localización b_k centrales (b_2, b_3 y b_4) correlacionaron con el puntaje promedio de los reactivos y sus coeficientes r resultaron entre -.95 y -.98, todos ellos significativos para un p < .001. La correlación más elevada se halló con el b_{prom} (r = -.99; p<.001) destacando que ambos indicadores reflejan una tendencia central del grado de adhesión que despertó el ítem. Los b_k extremos presentaron correlaciones más bajas (r_{b1} = -.72; p < .001; r_{b5} = -.67; n.s.) mostrando que estos parámetros tienen un comportamiento que no sigue el patrón establecido por la tendencia central. El sentido inverso de estas correlaciones también es esperable en tanto que, cuanto más nivel de confianza demande el ítem, menor cantidad de personas elegirán las categorías superiores del ítem y, consecuentemente, la media del reactivo tenderá a disminuir.

La distribución de los puntajes brutos de la prueba presentó una media de 28.6 (*DE* = 8.7) con un rango de variación entre 8 y 48 puntos. La misma resultó platicúrtica (*Cr* = -0.66.) y mostró un índice de asimetría negativa de -0.13. Como era de esperarse, dado el valor de estos índices, no se corroboró el ajuste al modelo normal con la Prueba de Kolmogorov-Smirnov (Z = 0.05; p<.0001). Se encontró una correlación de magnitud elevada de los puntajes brutos de la prueba con los θ estimados bajo el MRG (r = .99; p < .001). En el diagrama de dispersión de la figura 3 es posible apreciar la particularidad que tiene esta relación. La asociación resulta más estrecha para los niveles medios de la escala que para los niveles extremos.

El coeficiente alfa de Cronbach fue .90, lo que resulta bastante próximo al coeficiente de confiabilidad marginal. En efecto, ambos indicadores convergen en señalar que la prueba presenta una fiabilidad satisfactoria aun considerando el número reducido de ítems que la componen. También resultó elevado el glb con un valor de .93, confirmando una adecuada fiabilidad de la prueba.

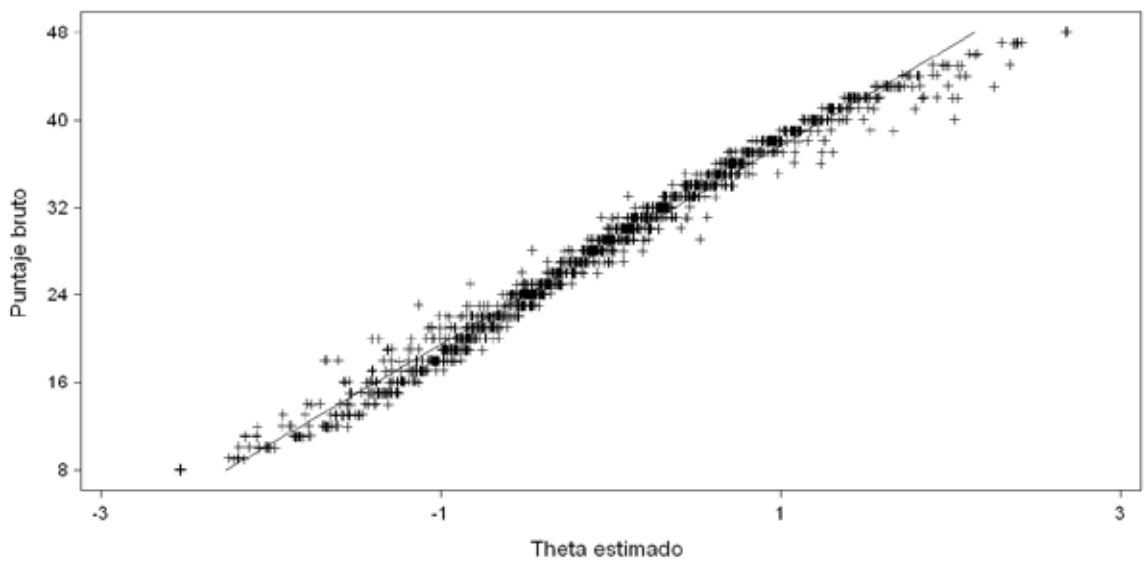


Figura 3.
Correlación entre puntaje bruto de la TCT y el theta desde la TRI.

Discusión

Los ocho elementos que conforman la escala de Confianza para la Matemática superaron las exigencias de calidad psicométrica de la TCT demostrando un comportamiento óptimo. Adicionalmente, la TRI también ofreció una valiosa comprensión del rasgo a partir de la interpretación de los parámetros de los ítems. El análisis de los b_{prom} y b_k reveló información sobre la relación existente entre el contenido del reactivo y la cantidad de Confianza que demanda no sólo el ítem sino, además, cada una de las categorías de la escala Likert. Esto resulta relativamente más sencillo en tests de rendimiento máximo, donde la complejidad creciente del ejercicio propuesto en el ítem se vincula a una demanda mayor de la habilidad medida. Sin embargo, en tests de comportamiento típico es mucho más arduo y delicado determinar cuánto rasgo se necesita para acordar o no con el contenido que propone el reactivo (Morizot et al. 2007).

El estudio del funcionamiento de las categorías de respuesta de los ítems mostró que la cantidad de opciones ofrece a los individuos una adecuada gama de matices para describir sus opiniones y todas resultan útiles para discriminar en algún rango de la variable. El hecho de que todos los ítems tengan un patrón similar de curvas podría ser atribuible a una propiedad de la escala Likert y a la manera en que los participantes interpretan los anclajes lingüísticos. Mientras que el resto de los umbrales inferiores fueron percibidos como aproximadamente equidistantes, la transición de la respuesta *En desacuerdo* hacia *Totalmente en desacuerdo* demandó un nivel de θ más alejado. Considerando que los ítems están redactados de forma inversa, podría suponerse que sólo aquellos presentaban una valoración extremadamente positiva de sus capacidades matemáticas tendieron a elegir *Totalmente en desacuerdo*.

Las relaciones encontradas entre las modelizaciones con la TCT y la TRI responden a lo esperable tanto desde una mirada teórica como empírica y son similares a las reportadas en otros trabajos (Asún & Zúñiga, 2008; Barbero et al., 2001). En efecto, Lord (1980) demostró que existe una relación monotónica entre los estadísticos de la TCT y los parámetros de la TRI. Ambas teorías persiguen objetivos idénticos: medir un rasgo latente y estimar el error inherente a este proceso. Para tal fin, estas teorías plantean un modelo y un conjunto de supuestos que, si se cumplen, garantizan la calidad de la medida (Muñiz, Fidalgo, García-Cueto, Martínez & Moreno, 2005). No se debe perder de vista que son construcciones que arrojan explicaciones plausibles con diferentes grados de complejidad y profundidad respecto de una misma realidad concreta (i.e. las respuestas observadas de los individuos a los ítems).

Una de las correlaciones más importantes remite a la encontrada entre θ y el puntaje bruto calculado según la TCT. Este hallazgo muestra una importante ventaja para la implementación práctica de la escala de Confianza para la Matemática. A la luz de la elevada correlación entre los puntajes de ambas teorías y empleando un principio de parsimonia, parece sensato recomendar para la mayoría de las aplicaciones un procedimiento de medición más simple como el que se propone desde la TCT. Esto evita la necesidad de requerir un programa específico que realice las estimaciones de los θ de los evaluados. La suma de las puntuaciones de cada uno de los ítems es una manera más sencilla de obtener una estimación del nivel de Confianza para la Matemática del individuo y no se diferencia significativamente de la obtenida según el MRG. Tampoco sería indispensable la estimación de θ al momento de buscar evidencias externas de validez, dado que no se reportan mayores beneficios (Ferrando & Chico, 2007).

No obstante, cabe aclarar que aunque se encontró una fuerte asociación lineal, el puntaje bruto es sólo una estimación de θ. La relación entre ambos es no-lineal porque θ se obtiene a partir de una transformación no-lineal del puntaje bruto (Lord, 1980). Puede apreciarse claramente en el diagrama de dispersión que representa la asociación de estos puntajes que la nube de puntos parece ajustarse a la forma de una función logística. Esto implica que para los niveles extremos del rasgo latente, el puntaje bruto registra un mínimo aumento del error en su predicción de θ.

Si las propiedades psicométricas y los puntajes obtenidos desde ambas teorías presentan elevadas correlaciones podría tornase aparentemente injustificable la aplicación de una metodología más costosa y sofisticada como la TRI (Kline, 2000). Sin embargo, cabe recordar que la TRI proporciona un status de rigurosidad y objetividad a la medición de constructos psicológicos imposible de alcanzar desde la perspectiva clásica. Al respecto, una de las ventajas de la TRI que ha sido corroborado empíricamente en este estudio es la propiedad de invarianza. Aún siendo estimados con muestras diferentes, se confirmó que los parámetros de los ítems se mantuvieron invariantes. Esto es diferente a lo que ocurre en la TCT, en donde las propiedades de los ítems dependen de la muestra de estandarización. Asimismo, se comprobó la invarianza de las mediciones respecto de la composición del instrumento. Las correlaciones alcanzadas entre los θ estimados a partir de dos

subconjuntos de cuatro ítems mostraron una intensidad elevada. Esto implica que, si los individuos responden cualquiera de los dos subconjuntos se podrían alcanzar niveles de rasgo que muestran una fuerte asociación.

Aunque se encontraron evidencias de un ajuste adecuado de los datos al MRG para todos los ítems de la prueba, conviene señalar que su evaluación se efectuó con los índices pobres que provee el MULTILOG (Revuelta et al., 2006). Este programa brinda proporciones observadas y esperadas para cada categoría de respuesta al ítem, pero la ausencia de discrepancia no implica que se prediga de manera adecuada la proporción de individuos que elige cada categoría en cada nivel del rasgo latente (Abad et al., 2011). La importancia del ajuste estadístico en el contexto de la TRI radica en su utilidad para controlar el error de medida en las respuestas (Van der Linden & Hambleton, 1997). La relativa novedad de los modelos politómicos origina que los métodos de evaluación de ajuste de los datos al modelo aún se encuentren en fase de exploración y redefinición. Para compensar los problemas en la evaluación del ajuste que presenta MULTILOG se obtuvieron evidencias indirectas de ajuste mediante otros indicadores surgidos del proceso de estimación de los parámetros (Rubio et al., 2007). En definitiva, los resultados obtenidos en calibración de los ítems dependen de cuán adecuado es el modelo para describir la respuesta de los individuos (Rojas & Pérez, 2001). Se trata de un problema de consistencia, si el modelo ajusta perfectamente eso quiere decir que se cumplen todas las condiciones requeridas por éste, se confirma la invarianza de las mediciones y de las poblaciones y se corrobora incluso la unidimensionalidad del constructo. En este sentido, los indicadores indirectos de ajuste analizados aportaron evidencias sobre esta consistencia.

Los estudios realizados desde la TCT y la TRI convergen en señalar que la confiabilidad de la prueba resultó satisfactoria. El coeficiente alfa de Cronbach, el glb y su análogo basado en la TRI, el coeficiente de confiabilidad marginal, tuvieron valores similarmente elevados. Además de estos resultados globales, la FIT demostró que el instrumento es preciso para medir en un rango bastante extenso de la variable. En consecuencia, en virtud de su relación inversa, el error típico de medida fue considerablemente bajo y parejo para una parte importante del espectro del rasgo latente por encima y por debajo de la media.

Los hallazgos obtenidos en el presente estudio constituyen un aporte sustancial de evidencias internas de validez y confiabilidad en el proceso de construcción de la escala de Confianza para la Matemática. La aplicación complementaria de la TCT y la TRI ha permitido explorar empíricamente las relaciones que pueden establecerse desde ambas perspectivas. Ulteriores investigaciones buscarán extender la muestra incorporando estudiantes pertenecientes a otras carreras humanístico-sociales. Un aumento en la heterogeneidad de la muestra permitiría obtener indicadores para estudiar la invarianza de los parámetros estimados en las distintas poblaciones dianas.

Referencias

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

Abal, F. (2013). *Comparación de modelos dicotómicos y politómicos de la Teoría de Respuesta al ítem aplicados a un test de comportamiento típico* (Tesis de Doctorado, Inédita, Universidad de Buenos Aires).

Abal, F., Lozzia, G. Aguerri, M., Galibert, M. & Attorresi, H. (2010). La escasa aplicación de la Teoría de Respuesta al Ítem en Tests de Ejecución Típica. *Revista Colombiana de Psicología, 19*(1) 111-122.

Asún, R. & Zúñiga, C. (2008). Ventajas de los Modelos Politómicos de la Teoría de Respuesta al Ítem en la Medición de Actitudes Sociales. El Análisis de un Caso. *Psykhe, 17*, 103 - 115.

Ayala, R. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

Baker, J. G., Rounds, J. B. & Zevon, M. A. (2000). A Comparison of Graded Response and Rasch Partial Credit Models with Subjective Well-Being. *Journal of Educational and Behavioral Statistics, 25,* 253-270.

Baker. F. B. (2001). *The Basics of Item Response Theory.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Barbero, M. I., Prieto, P., Suárez, J. C. & San Luis, C. (2001). Relaciones empíricas entre los estadísticos de la teoría clásica de los tests y los de la teoría de respuesta a los ítems. *Psicothema, 13*(2), 324-329.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord y M. R. Novick (Eds.). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Carmines, E. & Zeller, R. (1979). *Reliability and validity assessment*. Londres: Sage.

De Mars, C. (2010). *Item response theory*. New York: Oxford University Press.

Elosua, P. & Zumbo, B. D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema, 20*(4), 896-901.

Embretson, S. E. & Reise, S. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Erlbaum Publishers.

Ferrando, P. J. & Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT. *Psicológica, 28*, 237-257.

Gómez-Chacón, I. (2005). *Matemática emocional. Los afectos en el aprendizaje matemático*. España: Narcea.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. (1999). *Análisis Multivariante*. Madrid: Prentice Hall.

Hambleton R. K. & Swaminathan H. (1985). *Item Response Theory: Principles and applications.* Boston: Kluwer.

Hambleton, R. K. (2004). Theory, methods and practices in testing for the 21st century. *Psicothema*, *16*, 696-701.

Hernández, G. (2011). Estado del arte de creencias y actitudes hacia las matemáticas. *Cuadernos de Educación y Desarrollo, 3*(24). Recuperado el 10 de marzo de 2012, de http://www.eumed.net/rev/ced/index.htm.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood: Dow Jones-Irwin.

Jones, L. & Thissen, D. (2007). A History and Overview of Psychometrics. En C. R. Rao y S. Sinharay (Eds.), *Handbook of Statistics, 26: Psychometrics* (pp. 1–27). Amsterdam: North Holland.

Kaiser, H. F. (1960) The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

King, D. W., King, L. A., Fairbank, J. A., Schlenger, W. E. & Surface, C. R. (1993). Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory. *Psychological Assessment, 5*, 457-471.

Kline, P. (2000). *Handbook of Psychological Testing.* Londres: Routledge.

Kramp, U. (2006). *Efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad* (Tesis de Doctorado, Universidad de Barcelona). Recuperado de: http://www.tesisenred.net/bitstream/handle/10803/2535/UKD_TESIS.pdf?sequence=1

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

Lorenzo-Seva, U. & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*(1), 88-91.

Martínez-Arias, M. R. (1995). *Psicometría: Teoría de los Tests Psicológicos y Educativos.* Madrid: Síntesis.

Martínez-Padrón, O. J. (2008). Actitudes hacia la matemática. *Sapiens. Revista Universitaria de Investigación, 9*(1), 237-256.

Montero, I. & León, O. G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology, 7*(3), 847-862.

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. (1994). *Teoría Clásica de Test*. Madrid: Pirámide.

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

Muñiz, J., Fidalgo, A. M., García-Cueto, E., Martínez, R. & Moreno, R. (2005). *Análisis de los ítems*. Madrid: La Muralla.

Muraki, E. & Bock, R. (2003). *PARSCALE.* Chicago: Scientific Software International.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

Revuelta, J., Abad, F. J. & Ponsoda, V. (2006). *Modelos Politómicos de respuesta al ítem*. Madrid: La Muralla.

Rojas, A. J. & Pérez, C. (2001) *Nuevos Modelos para la Medición de Actitudes*. Valencia: Promolibro.

Rubio, V. J., Aguado, D., Hontangas, P. M. & Hernández, J. M. (2007). Psychometric Properties of an Emotional Adjustment Measure. An Application of the Graded Response Model. *European Journal of Psychological Assessment*, *23*(1), 39-46.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

Samejima, F. (1997). Graded Response Model. En W. J. Van der Linden. & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, (pp. 85-100). New York: Springer.

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107–120.

Spearman, C. E. (1904). General Intelligence objetively determined and measured. *American Journal of Psychology, 15*, 201 -293.

Tapia, M. & Marsh, G. E. (2004). An instrument to measure mathematics attitudes. *Academic Exchange Quarterly, 8*, 16-21.

Ten Berge, J.M.F., Snijders, T.A.B. & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis. *Psychometrika, 46*, 201-213.

Thissen, D. (2003). *MULTILOG*. Chicago: Scientific Software International.

Thurstone, L. L. (1928). Attitudes can be measured? *American Journal of Sociology, 33*, 529-554.

Timmerman, M. E. & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*, 209-20.

Van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

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