Revista de Psicología
UNIVERSIDAD DE CHILE

2016, *25*(2), 1-16
ISSN impreso: 0716-8039
ISSN en línea: 0719-0581
www.revistapsicologia.uchile.cl

Análisis psicométricos del Subtest de Razonamiento Numérico utilizando el Modelo de Rasch

Psychometric Analysis of Numerical Reasoning Subtest Using the Rasch Model

Marcos Cupani & Franco D. Cortez

Cipsi - Grupo Vinculado Centro de Investigaciones y Estudios sobre Cultura y Sociedad (CIECS)-Conicet, Universidad Nacional de Córdoba, Córdoba, Argentina

Resumen: El objetivo de este estudio fue evaluar las propiedades psicométricas del Subtest de Razonamiento Numérico del Test de Aptitudes Diferenciales, mediante el Modelo de Rasch. Se administró esta prueba a una muestra de 1.484 adolescentes de ambos sexos, con edades comprendidas entre 12 y 17 años (M = 14,00, DT = 1,38). Se evaluó la unidimensionalidad del instrumento, el ajuste de los ítems al modelo, los índices de separación y de fiabilidad para personas e ítems, el funcionamiento diferencial del ítem y la objetividad específica. De los 40 ítems, 38 presentaron un ajuste adecuado al modelo. El análisis de ajuste de las personas refleja que el 92% de los patrones de respuesta se ajustaron al modelo. Los índices de separación de los ítems (18,20) y de las personas (1,95), y los índices de fiabilidad de los ítems (1,00), así como los índices de fiabilidad de las personas (0,79), fueron satisfactorios. Dos ítems presentaron un comportamiento diferencial según el sexo. Los resultados son satisfactorios y contribuyen a esclarecer cómo este modelo psicométrico permite asegurar que los parámetros de las personas y de los ítems se expresen en las mismas unidades de medición, respetando las propiedades de intervalo.

Palabras clave: Modelo de Rasch, razonamiento numérico, unidimensionalidad, funcionamiento diferencial del ítem, objetividad específica.

Abstract: The aim of this study was to evaluate the psychometric properties of numerical reasoning subtest of the Differential Aptitude Test applying the Rasch model. This test was administered to a sample of 1,484 adolescents of both sexes, aged between 12 and 17 years old (M = 14.00; SD = 1.38). The unidimensionality of the instrument, adjustment of items to the model, separation rates and reliability for persons and items, item functioning differential (DIF), and the specific objective were evaluated. Of the 40 items, 38 showed an adequate fit to the model. The fitting analysis of people reflects that 92% of the response patterns were fitted to the model. Separation rates item (18.20), item reliability (1.00), and the values of individual separation (1.95) and reliability person (0.79) were satisfactory. Two items showed a differential behavior by sex. The results are satisfactory and help to clarify how this psychometric model allows parameters to ensure that persons and items are expressed in the same units of measurement respecting the properties of interval.

Keywords: Rasch model, numerical reasoning, unidimensionality, item functioning differential, specific objective.

Este trabajo fue financiado por la Secretaría Nacional de Ciencia y Tecnología (Foncyt-PICT12-994) y la Secretaría de Ciencia y Tecnología - Universidad Nacional de Córdoba (SECyT-UNC- RR Nº 1565-2014).

Contacto: M. Cupani. Investigador adjunto. Cipsi. Grupo Vinculado Centro de Investigaciones y Estudios sobre Cultura y Sociedad (CIECS)-Conicet, Facultad de Psicología, Universidad Nacional de Córdoba, Ciudad Universitaria, Córdoba 5000, Argentina. Correo electrónico marcoscup@gmail.com

Cómo citar: Cupani, M. & Cortez, F. D. (2016). Análisis psicométricos del Subtest de Razonamiento Numérico utilizando el Modelo de Rasch. *Revista de Psicología*, *25*(2), 1-16.
http://dx.doi.org/10.5354/0719-0581.2016.44558

Introducción

En las últimas evaluaciones del Programa para la Evaluación Internacional de Alumnos (PISA, por su nombre en inglés), cuyo objetivo es evaluar el grado de conocimiento que poseen los alumnos en áreas como ciencia, lectura y matemática, se ha observado que el rendimiento de los estudiantes secundarios argentinos en matemática fue desalentador (Organization for Economic Cooperation & Development, 2006): de un total de 57 países participantes, Argentina quedó ubicada en el puesto número 52.

En nuestro contexto, diferentes estudios se han enfocado en comprender este rendimiento y se ha evidenciado que para alcanzar un desempeño satisfactorio los estudiantes requieren adquirir un sentido de eficacia personal (Cupani & Lorenzo, 2010; Cupani & Zalazar Jaime, 2014; Zalazar Jaime, Cupani, & De Mier, 2015), poseer habilidades para organizar su trabajo, proponerse metas, monitorear sus progresos (Cupani & Pautassi, 2013) como también poseer aptitudes específicas (Cupani & Pautassi, 2013). Estos trabajos han corroborado el peso relativo de la variable aptitudes en relación con otras variables psicológicas (autoeficacia, expectativas y rasgos de personalidad) para el rendimiento académico.

Las aptitudes en matemáticas han sido operacionalizadas en Argentina mediante el Subtest de Razonamiento Numérico del Test de Aptitudes Diferenciales (DAT-5, por su nombre en inglés: Differential Aptitude Test) (Bennett, Wesman, & Seashore, 2000). El Subtest de Razonamiento Numérico mide la capacidad para comprender relaciones y conceptos expresados con números (multiplicaciones, divisiones, fracciones, relación, entre otros) y valora la capacidad de razonamiento más que la del cálculo. Este subtest ha demostrado poseer propiedades psicométricas adecuadas y ha sido utilizado en diferentes contextos y áreas de aplicación de la psicología (por ejemplo, la orientación vocacional).

Como regla general, la teoría psicométrica que se ha empleado para la construcción de los subtest del DAT-5 ha sido la teoría clásica de los test (TCT). Las características métricas de esta teoría, sin embargo, presentan un problema de doble invariancia: 1) las medidas para cada persona dependen del instrumento utilizado (por ejemplo, una misma persona obtendrá puntuaciones diferentes en el DAT-5, según la prueba utilizada); 2) las estimaciones de los ítems y las propiedades del test dependen de la muestra de individuos utilizada para este fin (por ejemplo, la fiabilidad de una prueba dependerá de la muestra utilizada para calcularla). Además, la TCT supone que una vez que la fiabilidad de una prueba ha sido estimada para una cierta población, esta precisión se mantiene constante para todos los niveles (es decir, seguirá siendo idéntica al estimar las medidas de las personas con valores bajos, medios y altos en aptitudes).

Los avances en psicometría han ayudado a sustituir la TCT en favor del uso de modelos basados en la teoría de respuesta al ítem (TRI). Con estos modelos se pueden obtener medidas invariantes, independientemente de los instrumentos utilizados y de los individuos evaluados (Engelhard Jr., 2013). El procedimiento de calibración es independiente de la muestra a la que se administra la prueba (es invariante sobre la población) y las medidas de las personas también están libres de la prueba que se aplique (no importa qué selección de ítems se utilice para estimar estos parámetros). Otras ventajas son que la TRI

permite calcular el error de medida para cada ítem y para cada persona, y por eso, en lo que respecta a la TRI, no tiene sentido referirse a la fiabilidad de la prueba como un atributo inamovible, ya que en este modelo la precisión de la medida (error de medida estándar) se estima para cada nivel de habilidad en la variable. Es decir, en este modelo se pone especial énfasis en el análisis de los ítems y en el nivel de habilidad de las personas a la hora de emitir una respuesta a los mismos. Es por eso que en la actualidad la TRI y el Modelo de Rasch, en particular, están ganando popularidad por encima de la TCT como modelos para crear y validar instrumentos (Embretson & Reise, 2000).

El Modelo de Rasch proporciona una metodología completa y detallada que permite evaluar las propiedades psicométricas de un instrumento a nivel de los ítems en función de las propiedades particulares de cada ítem que componen el test (Messick, 1994). Esto es, las puntuaciones que se obtienen de las pruebas vienen dadas en función de los ítems y de las personas que contestan los mismos, por lo tanto, el objeto de validez no es el test en sí mismo, sino la interpretaciones que se realizan (Messick, 1989).

El análisis de Rasch proporciona más información acerca de la capacidad de una persona porque se centra en la dificultad de los ítems, más que en la cantidad de ítems que responde correctamente cada participante. Desde la perspectiva de Rasch, la habilidad de una persona interactúa con la dificultad del ítem para poder así obtener una puntuación para cada sujeto en la medida (Linacre, 2002). El modelo propuesto por Rasch (1960) se fundamenta en el supuesto de que a) el atributo que se desea medir puede representarse en una única dimensión en la que se situarían conjuntamente las personas y los ítems, y b) el nivel de la persona en el atributo y la dificultad del ítem determinan la probabilidad de que la respuesta sea correcta. Rasch (1977) usó la función logística para modelar la relación:

$$p\,(x_j = 1|\theta, \alpha, \delta_j) = \frac{1}{1 + e^{-\alpha(\theta - \delta_j)}}$$

donde $p\,(x_j = 1|\theta, \alpha, \delta_j)$ es la probabilidad de responder con 1 (i.e., $x_j = 1$), θ es la ubicación de la persona y δj es el nivel de dificultad del ítem j. Expresado en palabras, la ecuación indica que la probabilidad de una respuesta correcta 1 es una función de la diferencia en el atributo entre el nivel de la persona (θ) y el nivel de dificultad del ítem (δj). Así, cuando una persona responde a un ítem equivalente a su umbral de competencia, tendrá la misma probabilidad de una respuesta correcta y de una respuesta incorrecta. En este caso, la dificultad del ítem es equivalente al nivel de competencia de la persona ($\theta_s - \delta j = 0$). Si la competencia del sujeto es mayor que la requerida por el ítem ($\theta_s - \delta j > 0$), la probabilidad de una respuesta correcta será mayor que la de una respuesta incorrecta. Por el contrario, si la competencia del sujeto es menor que la requerida por el ítem ($\theta_s - \delta j < 0$), la probabilidad de una respuesta correcta será menor que la de una respuesta incorrecta.

El Modelo de Rasch requiere que los ítems tengan un valor constante del parámetro de discriminación (α). Para el Modelo de Rasch, $\alpha = 1$ es igual para todos los ítems. Sin embargo, para algunos, el Modelo de Rasch representa una perspectiva filosófica diferente de la contenida en el modelo 1PL. El modelo 1PL (como también los de dos y tres parámetros) se

enfoca en establecer el mejor ajuste de los datos. Por el contrario, se considera que el Modelo de Rasch es el estándar para la construcción de un instrumento para medir la variable de interés (Wilson, 2005) y que se enfoca en establecer el grado en que los datos se ajustan a este modelo de medición.

Para analizar los ítems, el Modelo de Rasch primero convierte los datos ordinales de un instrumento en datos intervalares, cumpliendo de este modo con uno de los requisitos indispensables de toda medición (Kleinman & Teresi, 2016). Luego, este modelo psicométrico permite evaluar varias características, como el nivel de ajuste del modelo, la dificultad y el orden jerárquico de los ítems, la fiabilidad de las personas e ítem, los índices de separación y el funcionamiento diferencial del ítem (DIF, por su nombre en inglés).

Para evaluar estas características debe tenerse en cuenta que los datos empíricos se ajusten al modelo propuesto (Prieto & Delgado, 1999). El ajuste del ítem se refiere a cuán bien cada reactivo mide el constructo de interés (Bond & Fox, 2003) y se cuantifica mediante medidas de infit y outfit, lo que permite asegurar que el instrumento evalúe de forma correcta el constructo que pretende. El grado de acuerdo entre el patrón de respuestas observadas y las expectativas del modelo son establecidos por los estadísticos de ajuste. Los índices de ajuste ayudan a determinar si los parámetros estimados de los ítems pueden ser considerados como un resumen del patrón de repuesta observado. La jerarquización de los ítems consiste en el ordenamiento de los mismos en niveles de dificultad (del más fácil al más difícil). Este ordenamiento de ítems es un principio fundamental de la edición dado que nos permite determinar si un participante posee mayor o menor habilidad con respecto a otro (Bond & Fox, 2003). Si el ítem no está en escala logit y ordenado jerárquicamente, las puntuaciones obtenidas por un participante en el test pueden ser engañosas. El ordenamiento jerárquico de los ítems nos permite identificar ítems redundantes o niveles de dificultad no cubiertos que disminuyen la precisión y la eficacia del instrumento.

La precisión de la medida depende de cuán bien los ítems del instrumento permitan diferenciar los niveles de habilidad. Los índices de separación de personas son una estimación de cuán bien el instrumento puede diferenciar a las personas en la medida. El análisis de Rasch ofrece estadísticas de fiabilidad y separación para los ítems y las personas. La fiabilidad significa el grado de reproducibilidad de las habilidades relativas o de las dificultades estimadas (Linacre, 2016). Es decir, un índice alto en fiabilidad para personas nos indica que existe una alta probabilidad de que las personas identificadas por el test con alta habilidad posean realmente esas habilidades y no otras. De manera semejante, alta fiabilidad en los ítems significa que los ítems establecidos como de alta dificultad tienen realmente alta dificultad y no baja dificultad. El índice de separación indica el número de diferentes estratos de rendimiento que la prueba puede identificar (Wright, 1996).

El DIF puede ser conceptualizado como el hecho de que la respuesta a un ítem está sujeta a cambios en función de diferentes grupos de personas (De Ayala, 2009). En otras palabras, un ítem presenta DIF cuando la probabilidad de respuesta correcta no depende únicamente del nivel

de la persona en el rasgo intencionadamente medido por el test (Bond & Fox, 2003).

En resumen, estas propiedades de análisis de Rasch pueden ayudar a investigadores y educadores a mejorar la validez, la fiabilidad y la eficiencia de los instrumentos de medición (Bond, 2003). Por lo tanto, el propósito de este trabajo es evaluar las propiedades psicométricas del Subtest de Razonamiento Numérico del DAT-5 (Bennett et al., 2000) mediante el Modelo de Rasch y demostrar sus ventajas tanto metodológicas como aplicadas.

Método

Participantes

La muestra estuvo compuesta por 1.484 adolescentes de ambos sexos, 807 mujeres (54,6%) y 665 hombres (44,8%); que cursaban primero y segundo año del ciclo final de Educación General Básica (EGB), con edades comprendidas entre 12 y 17 años (M = 14,00; DT = 1,38) y que realizaban sus estudios en colegios estatales (43%) y privados (57%) de la ciudad de Córdoba, Argentina.

Instrumento

Subtest de Razonamiento Numérico. Mide la capacidad para comprender relaciones y conceptos expresados con números. Está compuesta por 40 preguntas de opciones múltiples, con cinco alternativas de respuesta de las cuales una es correcta y las restantes son distractores. Este subtest pertenece al DAT-5, que mide ocho aptitudes: razonamiento verbal, razonamiento numérico, razonamiento abstracto, rapidez y exactitud perceptiva, razonamiento mecánico, relaciones espaciales, ortografía, y uso del lenguaje. Bennett et al. (2000) reportan índices de fiabilidad adecuados de los ocho subtest (valores de KR-20 entre ,75 y ,92).

Procedimiento

Para este trabajo se contó con los datos ya recogidos entre los años 2007 y 2012 en el proyecto de investigación titulado “El rol de la personalidad en un modelo social-cognitivo de rendimiento académico” y subsidiado por el Fondo para la Investigación Científica y Tecnológica de la Agencia Nacional de Promoción Científica y Tecnológica. Para ese proyecto el Subtest Razonamiento Numérico fue administrado por uno de los autores de este trabajo con la colaboración de estudiantes de la Facultad de Psicología de la Universidad Nacional de Córdoba. La administración fue colectiva y en un horario regular de clases, con autorización previa de los profesores de cada curso; se solicitó la colaboración de cada alumno y se enfatizó la naturaleza voluntaria y anónima de su participación. Se utilizaron formularios de consentimiento informado y se tomaron medidas para garantizar el respeto de los derechos humanos y el cuidado del medioambiente. Además, se realizó un estricto control para evitar cualquier riesgo emergente, y para garantizar el buen uso y manejo de la información. Los investigadores de este proyecto declaran conocer y realizar las salvaguardas previstas en la Declaración de Helsinki, así como la ley 25.326 de Principios generales relativos a la protección de datos, aprobada por el Congreso Argentino el 04/10/2000.

Análisis de datos

Todos los análisis se realizaron con el Modelo de Rasch, que consigue que todos

los parámetros de las personas (θ) y los ítems (δ) sean localizaciones puntuales en una única variable latente, que pueden ser expresadas en la misma unidad de escala (logit) y que posibilitan establecer comparaciones objetivas. El plan de calibración del Subtest de Razonamiento Numérico del DAT-5 consistió en los siguientes pasos.

Paso a. Unidimensionalidad e independencia local. La unidimensionalidad se evaluó con el Método Robusto para el Análisis Armónico de la Ojiva Normal (NOHARM, por su nombre en inglés Normal Ogive Harmonic Analysis Robust Method) mediante el programa NOHARM versión 4.0, que permite evaluar la relación entre el análisis factorial no-lineal y el modelo de ojiva normal en orden del ajuste unidimensional y multidimensional del modelo de ojiva normal (De Ayala, 2009). NOHARM produce una matriz residual para evaluar el ajuste del modelo. Esta matriz residual es la discrepancia entre la matriz de covarianza observada y la matriz de covarianza de los ítems luego de ajustar el modelo. El software provee la raíz de la media de los residuos al cuadrado (RMSR, por su nombre en inglés root mean square of residuals), donde valores cercanos a 0 representan un adecuado ajuste al modelo. Si el RMSR es superior al error típico de los residuos (que es $4\sqrt{N}$) indica que el modelo no se ajusta bien. Una segunda medida de ajuste es el índice de Tanaka (1993) de bondad de ajuste (GFI, por su nombre en inglés Goodness of Fit Index). McDonald (1989) sugiere que un puntaje de ,90 es un valor aceptable, un índice de ,95 indica un buen ajuste y un valor igual a 1 indicaría un ajuste perfecto. El supuesto de independencia local se evaluó inspeccionando la matriz de los residuos (se esperan valores inferiores a 0,025) y la matriz de varianza y covarianza (se esperan valores inferiores a 0,25). De manera complementaria, la unidimensionalidad de la escala se evaluó mediante el análisis de componentes principales de los residuos (PCAR, por su nombre en inglés Principal Components Analysis of Residuals). Se consideró que se cumple el supuesto de la unidimensionalidad si el modelo de medida (el Modelo de Rasch unidimensional) explicaba aproximadamente un 50% de la varianza. Si el mayor factor adicional (una dimensión secundaria) tiene un valor propio menor de tres (una fuerza de tres ítems) y explica menos del 5% de la varianza inexplicada, la unidimensionalidad puede sostenerse (Linacre, 2016).

Paso b. Ajuste del Modelo de Rasch. Se utilizó el algoritmo joint maximum likelihood para observar si los datos se ajustan al Modelo de Rasch, y se realizaron tres análisis: el ajuste global de los datos, el ajuste de los ítems y el ajuste de las personas. Con el primero se comprueba si, en términos generales, la matriz de datos se ajusta a lo pronosticado por el modelo. El ajuste de los ítems permite estudiar cada uno de estos de manera independiente. Asimismo, con el ajuste de las personas se identifican a las personas que han respondido de manera incoherente a la formulación teórica. Se utilizaron dos estadísticos que aportan información sobre el ajuste de los datos al modelo: el Infit MS, que es un índice de ajuste interno que evalúa el ajuste entre los parámetros próximos entre sí, y el Outfit MS, que es un índice de ajuste externo que evalúa el ajuste con respecto a parámetros alejados. Cuando los datos observados coinciden con los propuestos por el modelo, los valores de Infit MS y Outfit MS tienen valores próximos a 1; en caso contrario, se obtendrán valores alejados de 1. Es decir, un valor de Infit MS de 1 indica

que el 100 % de la varianza de los datos empíricos son explicados por el modelo, mientras que un valor de 1,3 indica que hay más varianza de la esperada (un 30% de la varianza no puede ser explicada por el modelo).

Siguiendo los criterios propuesto por Linacre (2002) la región para considerar un ajuste aceptable oscila entre 0,6 y 1,3. También se calcularon los coeficientes biserial puntual (rpbis, por su nombre en inglés point-biserial correlation discrimination estimates), que son un indicador útil para diagnosticar errores en la codificación de ítems o de claves (valores negativos o en 0 indican ítems o personas con patrones de respuesta que contradicen la variable).

Paso c. Separación y fiabilidad. Los ítems deben estar suficientemente bien separados en niveles de dificultad para poder identificar el sentido y el significado de la variable latente (Wright & Stone, 2003). El índice de separación de las personas indica cuán bien el instrumento puede discriminar a estas sobre la variable medida. Un conjunto útil de ítems debe definir al menos tres estratos de personas (por ejemplo, los niveles altos, moderados y bajos de conocimiento). El índice de separación superior a 2 se considera adecuado como también una fiabilidad asociada al índice de separación de 0,80 (Bond & Fox, 2003).

Paso d. DIF. Se realizaron análisis de DIF según el sexo de los participantes. Para aplicar el DIF se realizaron análisis pairwise en donde el nivel de significación se fijó en $\alpha < ,01$, y se tuvo en cuenta que el contraste del DIF debe ser superior a $\geq ,5$ logits (Linacre, 2016). Para este análisis, Winsteps utiliza la *t* de Welch (Linacre, 2016), que se obtiene al dividir el contraste DIF por el error estándar conjunto de las medidas DIF. El contraste DIF es la diferencia entre los tamaños DIF y sus estimaciones log-odds.

Paso e. Objetividad específica. Se realizó un análisis de la objetividad específica de los ítems (Rasch, 1977). Esta es una de las propiedades más importantes del Modelo de Rasch y hace referencia a que una medida solo puede ser considerada válida y generalizable si no depende de las condiciones específicas con las que ha sido obtenida. Uno de los principales procedimientos que se recomiendan para analizar el ajuste de los datos al modelo consiste en contrastar empíricamente esta propiedad. En este estudio, para analizar la invarianza de los parámetros de los ítems: i) se dividió la base en dos en forma aleatoria, ii) se estimaron los parámetros de dificultad de los ítems, y iii) se llevó a cabo una regresión lineal simple entre los parámetros obtenidos. Los valores esperados para la correlación entre ambos conjuntos de parámetros, la ordenada en el origen y la pendiente de la recta que indican un ajuste perfecto serían 1, 0 y 1 respectivamente (Prieto & Delgado, 1999).

Resultados

Paso a. El valor del RMSR (0,0094) es menor que el error típico de los residuos estimado (0,1038), lo que nos indica que se cumple el supuesto de unidimensionalidad. El índice de Tanaka de bondad de ajuste fue 0,96, lo que muestra un buen ajuste. Por otro lado, los valores de la matriz de varianza y covarianza no fueron superiores al valor de corte de 0,25. Además, se observó que 1,4% de los residuos de todos los ítems fueron inferiores a 0,025, lo que indica que se cumple el supuesto de independencia local. Por otro lado, el análisis de componentes principales de los residuos

mostró que la dimensión de Rasch explicó 47,1% de la varianza en los datos con su valor propio de 35,6. El primer contraste (la mayor dimensión secundaria) tuvo un valor propio de 2,4 y representó el 3,2% de la varianza no explicada. Estos resultados globales indican que se cumple el supuesto de unidimensionalidad.

Paso b. Centrados en el análisis de los ítems (ver tabla 1), el ajuste ha resultado satisfactorio para 38 de los 40 ítems que forman el DAT-5 (Ítem Infit MS y Outfit MS $\leq$ 1,3). Los valores de Outfit MS obtenidos para dos ítems (35-Divisor y 39-Algoritmo) que no se ajustan están indicando un comportamiento poco predecible por el modelo (Linacre, 2002).

Además, esos mismos ítems presentaron valores de rpbis cercanos a cero. La medida de dificultad (δ_i) de los ítems varió entre $-3{,}50 \leq \delta_i \leq 1{,}77$, con una media de 0,00 (*DT* = 1,18). Los valores de Infit de los ítems variaron entre 0,89 y 1,13, con una media de 1,00 (*DT* = 0,06), y los índices de Oufit entre 0,84 y 1,41, con una media de 1,03 (*DT* = 0,13). El análisis de ajuste de las personas refleja que el 92 % de los patrones de respuesta se ajustaron al modelo (Infit y Outfit $\leq$ 1,3). Los niveles de habilidad variaron entre $-3{,}61 \leq \theta \leq 2{,}37$ con una media de -0,34 (*DT* = 0,83).

En el mapa de personas e ítems, también conocido como Mapa de Wright, ilustrado en la figura 1, se muestra la distribución contigua de las personas y de los ítems de manera conjunta. Se puede observar en el lado izquierdo la distribución de los niveles de habilidad de las personas de nuestro estudio y en el lado derecho la dificultad de los ítems.

En el gráfico se puede observar que la mayoría de los ítems se ubican en una posición centrada con respecto a los estudiantes evaluados y que los ítems, en líneas generales, logran una adecuada distribución por el continuo, aunque podría ser necesario añadir algunos ítems para cubrir el sector medio del continuo (entre los ítems 3 y 8). Del mismo modo, se puede observar que algunos ítems (por ejemplo, 6, 14, 18 y 33) son redundantes en cuanto a su dificultad (δ entre 0,11 y 0,18) y que miden el mismo contenido (Ecuación). Los datos también indican que el test es levemente difícil (media de δ = 0,00) para la muestra de estudiantes analizada (media de θ = -0,34), por lo tanto, se podría pensar que faltarían algunos ítems de baja dificultad para evaluar adecuadamente a los sujetos con baja competencia. En efecto, existe un rango aproximadamente de 1,37 logit entre el ítem 1 (-3,50) y el ítem 9 (-2,13), que son los dos ítems fáciles de la prueba. Sin embargo, la zona de alineamiento entre la dificultad de los ítems y la presencia del rasgo latente en los sujetos agrupa a 1.436 estudiantes (96,77%). Con relación al contenido medido por los ítems, se puede observar que siete de los 10 ítems de menor dificultad hacen referencia a contenidos relacionados con ecuación; mientras que de los 10 ítems de mayor dificultad, cinco están relacionados con división y dos con lógica, aunque los dos ítems de mayor dificultad del test son sobre contenidos relacionados con las ecuaciones.

Tabla 1
Características de los 40 ítems del Subtest de Razonamiento Numérico del DAT-5

Ítem	Parámetros de Rasch				DIF		
	δ_i	SE	Infit MS	Outfit MS	Rasch-Welch (*t*)	Femenino	Masculino
1. Lógica	-3,50	0,12	1,00	1,13	-0,68	-3,63	-3,47
2. Ecuación	-0,83	0,06	1,03	1,03	-1,93	-0,93	-0,71
3. Resta	-0,22	0,06	0,98	0,98	0,55	-0,19	-0,25
4. Ecuación	-1,73	0,07	1,03	1,04	-1,16	-1,81	-1,66
5. Redondeo	-0,64	0,06	1,03	1,03	0,54	-0,62	-0,68
6. Ecuación	0,17	0,06	0,97	0,98	-0,17	0,18	0,20
7. Ecuación	-0,69	0,06	0,96	0,93	-1,03	-0,74	-0,62
8. Divisor	0,06	0,06	1,00	0,98	-1,70	-0,03	0,16
9. Ecuación	-2,13	0,07	0,98	0,97	-1,17	-2,22	-2,04
10. División	-0,49	0,06	1,11	1,16	2,62	-0,35	-0,65
11. División	0,52	0,06	1,07	1,14	-0,22	0,53	0,56
12. Ecuación	-1,14	0,06	1,03	1,04	0,91	-1,09	-1,20
13. Ecuación	0,20	0,06	0,91	0,90	-0,85	0,17	0,26
14. Ecuación	0,11	0,06	0,99	0,97	2,25	0,23	-0,03
15. Lógica	-1,62	0,06	0,99	0,99	-1,36	-1,71	-1,53
16. Ecuación	-1,43	0,06	0,96	0,97	1,74	-1,33	-1,54
17. Ecuación	-2,04	0,07	0,92	0,86	0,75	-2,00	-2,11
18. Ecuación	0,18	0,06	0,93	0,91	-3,48	0,00	0,40
19. Ecuación	0,65	0,06	0,92	0,87	-1,27	0,58	0,74
20. Ecuación	-0,61	0,06	0,95	0,92	-4,20	-0,83	-0,35
21. División	1,09	0,07	1,07	1,15	2,26	1,23	0,93
22. Divisor	-0,33	0,06	0,96	0,95	-3,83	-0,53	-0,10
23. Ecuación	0,30	0,06	1,02	1,01	-1,96	0,20	0,43
24. Lógica	1,02	0,06	1,04	1,06	0,56	1,06	0,99
25. División	-0,33	0,06	0,98	0,99	-1,08	-0,39	-0,27
26. División	-0,38	0,06	0,91	0,89	-3,76	-0,57	-0,14
27. División	0,98	0,06	1,05	1,10	-1,31	0,90	1,07
28. Ecuación	0,50	0,06	0,96	0,94	-2,17	0,38	0,65
29. Ecuación	1,77	0,08	1,04	1,19	2,36	1,97	1,59
30. Ecuación	0,50	0,06	0,97	0,97	5,96	0,86	0,14
31. Lógica	-0,49	0,06	0,92	0,90	3,63	-0,30	-0,71
32. Ecuación	0,44	0,06	0,89	0,85	-0,74	0,40	0,48
33. Ecuación	0,14	0,06	0,97	0,97	-0,21	0,13	0,15
34. Ecuación	0,86	0,06	0,97	0,98	2,57	1,01	0,69
35. Divisor	1,68	0,08	1,08	1,29	3,58	1,94	1,38
36. División	1,71	0,08	1,13	1,41	0,75	1,77	1,65
37. Ecuación	1,73	0,08	1,06	1,12	3,17	1,99	1,49
38. Divisor	1,00	0,06	1,07	1,13	1,06	1,06	0,92
39. Algoritmo	1,61	0,08	1,10	1,37	2,17	1,76	1,43
40. Algoritmo	1,38	0,07	1,02	1,05	0,18	1,40	1,37

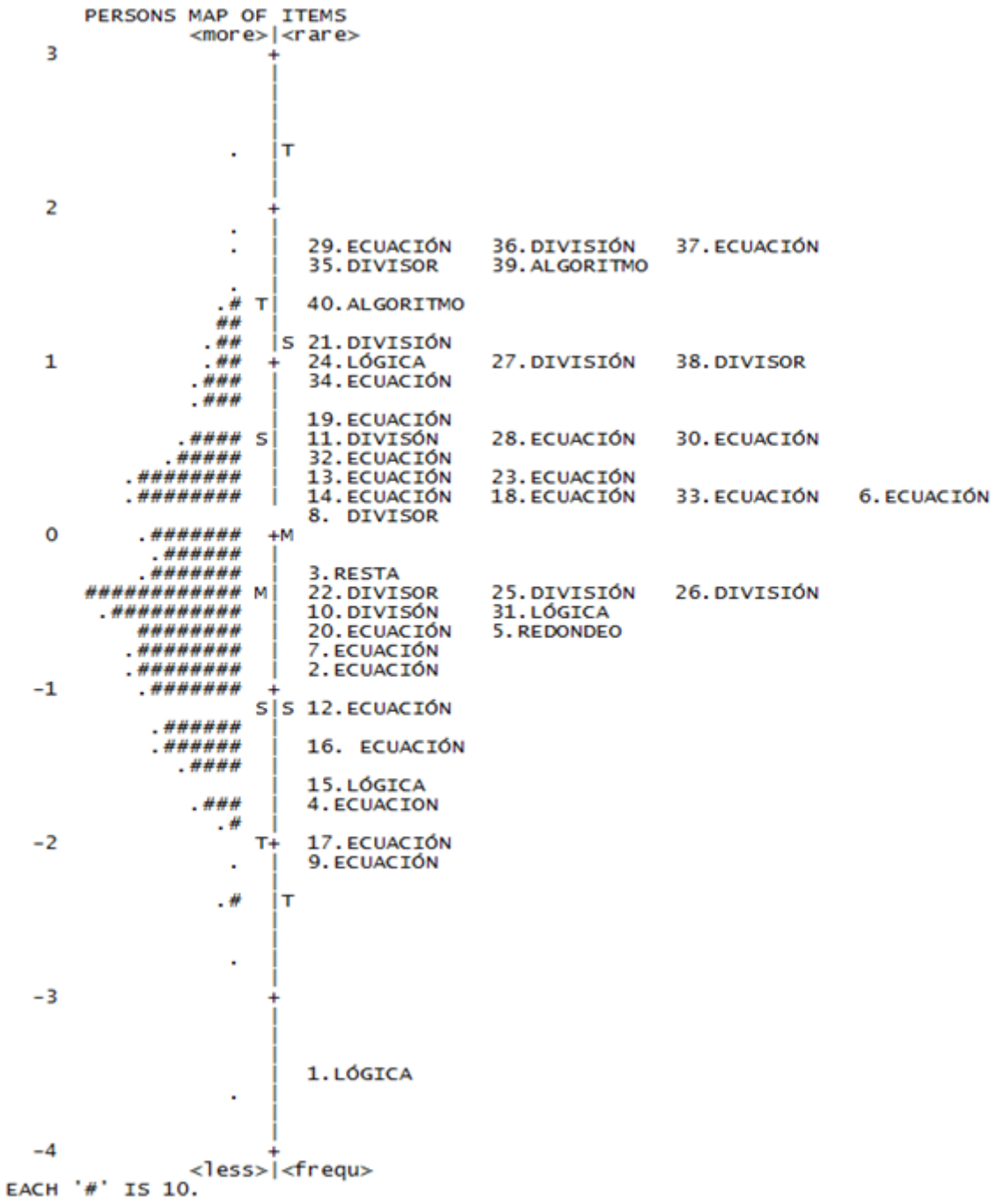

Figura 1. Mapa de personas e ítems. En la columna de la izquierda se observa la ubicación de las personas en el continuo según su nivel de habilidad. El símbolo # representa un grupo de cinco personas y el símbolo "." representa grupos de una a cuatro personas. Esta distribución suele asumir una forma de curva normal. M marca la media de las personas y los ítems. S es una DT alejada de la media. T es dos DT alejadas de la media.

Paso c. Los índices de separación de los ítems (18,20) y los índices de fiabilidad de los ítems (1,00) fueron satisfactorios, lo que nos indica que la muestra utilizada es suficientemente grande como para confirmar la jerarquía de dificultad del ítem (validez de constructo) del instrumento (Linacre, 2016). Por otro lado, los índices de separación de las personas (1,95) y los índices de fiabilidad de las personas (0,79) fueron considerados aceptables, aunque se puede considerar la necesidad de cubrir algunos niveles de habilidad con otras preguntas, ya que este pool de ítems puede no ser suficientemente sensible para distinguir entre sujetos de alto y de bajo rendimiento.

Paso d. Los resultados del análisis de DIF según el género permiten observar que el contraste DIF en el ítem 35 (Divisor) e ítem 37 (Ecuación) fue estadísticamente significativo (p < ,01), con un contraste del DIF de 0,56 y 0,51 respectivamente. La dificultad de los ítems (media de DIF) para la muestra masculina fue 1,38 (ítem 35) y 1,49 (ítem 37) logits mientras que para las mujeres fue 1,94 (ítem 35) y 1,99 (ítem 37). Esto indica que estas preguntas son más difíciles para las mujeres.

Paso e. Los resultados mostraron un valor de r = 0,994, el valor de la constante fue 0,001 y β = 0,994, por lo que se puede asumir la invariancia de los parámetros de los ítems anclas (Prieto & Delgado, 1999).

Discusión

El objetivo de este trabajo fue examinar las propiedades psicométricas del Subtest de Razonamiento Numérico mediante el Modelo de Rasch (Rasch, 1960). Este modelo psicométrico permite asegurar que los parámetros de las personas y de los ítems se expresen en las mismas unidades (medición conjunta), ajustar los datos al modelo demostrando qué personas son independientes de los ítems administrados (objetividad específica), y que la escala presenta propiedades de intervalo (propiedades de medida) como es el tipo logit (Schulz & Fraillon, 2011).

En líneas generales, los ítems que componen el Subtest de Razonamiento Numérico presentaron propiedades psicométricas aceptables. Los índices de dificultad y los niveles de habilidad de los participantes cubrieron gran parte del continuo medido, y los índices de fiabilidad (personas e ítems) indican que la localización de las personas e ítems sería previsiblemente reproducible (Andrich, 2002). El ajuste global de los ítems fue adecuado para casi todos estos. La zona de alineamiento entre la dificultad de los ítems y la presencia del rasgo latente en los sujetos fue alta y se observó que un 92% de los participantes respondieron coherentemente los ítems del test, lo que permitió identificar patrones de respuestas predecibles por el modelo propuesto (Linacre, 2002). El estudio de invarianza corroboró que los parámetros obtenidos de los ítems en dos submuestras son semejantes.

Aunque en líneas generales el instrumento funciona adecuadamente para medir el constructo de interés, es necesario destacar que los análisis nos han permitido identificar que dos ítems (36 y 39) no presentaron un buen ajuste a los datos, y otros dos ítems (35 y 37) presentaron un comportamiento diferencial según el sexo de los estudiantes. Los autores no encuentran motivos teóricos que expliquen este deficiente comportamiento de estos cuatro ítems, aunque sí resulta interesante ver que precisamente estos ítems que desajustaron requerían una habilidad importante

para ser respondidos, y que algunas personas que acertaron pudieron haberlo logrado por mera coincidencia. En lo que se refiere al DIF, también se observó que las mujeres encuentran más dificultad que los varones en dos de los ítems más complejos de la prueba, pero el contraste de ambos no justificaría una reedición de estos ítems. Aunque algunos estudios han demostrado que los niños obtienen mejores puntuaciones en geometría y cálculo y que las niñas obtienen mejores puntuaciones en análisis de datos y álgebra (García, Tello, Abad, & Moscoso, 2007), en este caso, no existe evidencia de que esa sea la razón de los resultados, por lo que habría que considerar que esta diferencia en los ítems se trata simplemente de un dato más sin relevancia sustantiva (Andrich & Hagquist, 2012).

Los resultados de este trabajo tienen tanto implicancias metodológicas como prácticas. Las implicancias metodológicas giran alrededor del uso del análisis de Rasch como control de calidad para analizar, evaluar y validar los instrumentos de medición. Es decir, desde este enfoque se puede analizar la validez de constructo, ya que los ítems que componen un instrumento deben estar distribuidos de forma creciente en dificultad como también estar adecuadamente representados por los contenidos de interés. La ausencia de ítems que se corresponda con cada nivel de habilidad es un indicador de la necesidad de contar con un número mayor de ítems para lograr una mejor cobertura de la prueba. El Modelo de Rasch también puede proporcionar un análisis detallado de los patrones de respuesta individuales que reflejan los procesos de razonamiento de los individuos involucrados. Agregando un análisis sobre los distractores se podría determinar cómo los estudiantes entienden e interpretan la consigna del ítem y por qué eligen una opción con respecto a otra del test. Por otro lado, al cumplir con el supuesto de unidimensionalidad mediante el análisis factorial no lineal y también al obtener un ajuste satisfactorio del modelo, se genera evidencia de que los ítems miden el constructo de interés y de que también las personas poseen las habilidades medidas por el instrumento.

Considerando las implicaciones prácticas, estos ítems pueden ser cargados en programas especializados y, de esa manera, utilizar Test Adaptativos Informatizados, lo que propiciaría minimizar el error estándar de medición y la posibilidad de medidas de longitud sin pérdida de precisión y de fiabilidad (Abad, Olea, Aguado, Ponsoda, & Barrada, 2010). Asimismo, se mejoraría la posibilidad de diagnóstico con evaluaciones más breves y precisas (Olea, Abad, Ponsoda, & Ximénez, 2004). Este diagnóstico de las habilidades y el patrón de repuestas a los ítems que realiza un estudiante puede ser analizado desde un enfoque cualitativo, y de este modo establecer cuáles son las operaciones matemáticas que se necesita fortalecer (Long, Wendt, & Dunne, 2011). Finalmente, los investigadores relacionados con el campo de las matemáticas pueden utilizar con mayor confianza este subtest, ya que logró resistir los análisis propios del Modelo de Rasch, y también pueden seleccionar (según el nivel de dificultad) un conjunto menor de ítems para medir este constructo con casi la misma precisión.

En términos generales, los resultados de este estudio son satisfactorios, pues permiten someter los ítems de este subtest a un modelo psicométrico más riguroso. No obstante, existen limitaciones para considerar. Una de las limitaciones es que solo se analizó el Subtest de Razonamiento Numérico y no se efectuaron los mismos análisis al resto de los test de la batería.

Por ello se recomienda para próximos estudios el análisis de los otros subtest que componen la prueba, esto con el fin de poder calibrar los ítems y poder generar test adaptativos informatizados. Por otro lado, una limitación relacionada con el análisis de los distractores mediante el Modelo de Rasch nos permitiría establecer si todos los distractores fueron utilizados por los estudiantes y si algunos ítems presentaron un funcionamiento inesperado de los distractores (Hammouri & Sabah, 2010).

En futuras investigaciones se planifica poder contar con el software RUMM2030 (Andrich, Sheridan, & Luo, 2013) para realizar este tipo de análisis. Otra de las limitaciones es que los participantes fueron seleccionados mediante un procedimiento de muestreo accidental y el uso de una muestra no probabilística puede haber afectado la estimación de los parámetros de los ítems. Esta limitación, sin embargo, no es un factor determinante para el cálculo de los parámetros de los ítems y personas cuando se utilizan Modelos de Rasch, aunque este no es el caso de otros modelos de TRI. Además, la muestra no es representativa de las distintas orientaciones y especialidades académicas que poseen los colegios estatales y privados, los cuales no han sido cubiertos de un modo exhaustivo en este estudio, así quedaron excluidas orientaciones como carreras técnicas, por ejemplo, por lo que no está claro si estos resultados serían generalizables a otras muestras más diversas. En futuras investigaciones sería necesario replicar estos análisis con una muestra de estudiantes de orientación técnica, donde se puede considerar que las competencias en matemáticas son más elevadas. Finalmente, en este trabajo no se pudo registrar en todos los casos el desempeño académico de los estudiantes en la asignatura matemáticas, los que nos hubiese permitido evaluar la validez predictiva del instrumento y también realizar un estudio de funcionamiento diferencial de ítems considerando a los estudiantes de alto y bajo desempeños académicos. Se planifica solicitar a las instituciones participantes la facilitación de estos datos para complementar los estudios propuestos.

En síntesis, los resultados de este trabajo contribuyen a esclarecer la aplicación de este tipo de modelos psicométricos y la importancia de realizar estos análisis a diferentes pruebas, por lo que se sugiere que futuras investigaciones pongan el foco en la construcción y/o adaptación de pruebas utilizando el Modelo de Rasch.

Referencias

Abad, F. J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J. R. (2010). Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT. *Psicothema*, *22*(2), 340-347. Recuperado de http://goo.gl/fAjIj6

Andrich, D. (2002). Implications and applications of modern test theory in the context of outcomes based education. *Studies in Educational Evaluation*, *28*(2), 103-121. http://dx.doi.org/10.1016/S0191-491X (02)00015-9

Andrich, D. & Hagquist C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*(3), 387-416. http://dx.doi.org/10.3102/1076998611411913

Andrich, D., Sheridan B. S., & Luo G. (2013). RUMM2030: An MS Windows computer program for the analysis of data according to Rasch unidimensional models for measurement. Perth, Australia: RUMM Laboratory. http://www.rummlab.com/

Bennett, G. K., Wesman, A. G., & Seashore, H. G. (2000). *DAT-5: tests de aptitudes diferenciales (versión 5: manual)*. Madrid, España: TEA Ediciones.

Bond, T. G. (2003). Relationships between cognitive development and school achievement: a Rasch measurement approach. En R. F. Waugh (Ed.), *On the forefront of educational psychology* (pp. 37-46). New York, New York: Nova Science Publishers. Recuperado de http://goo.gl/44zcbG

Bond, T. G. & Fox, C. M. (2003). Applying the Rasch model: Fundamental measurement in the human sciences. *Journal of Educational Measurement*, *40*(2), 185-187. http://dx.doi.org/10.1111/j.1745-3984.2003.tb01103.x

Cupani, M. & Lorenzo, J. (2010). Evaluación de un modelo social-cognitivo del rendimiento en matemática en una población de preadolescentes argentinos. *Infancia y Aprendizaje*, *33*(1), 63-74. http://dx.doi.org/10.1174/021037010790317216

Cupani, M. & Pautassi, R. M. (2013). Predictive contribution of personality traits in a socio-cognitive model of academic performance in mathematics. *Journal of Career Assessment*, *21*(3), 395-413. http://dx.doi.org/10.1177/1069072712475177

Cupani, M. & Zalazar Jaime, M. F. (2014). Rasgos complejos que predicen el rendimiento académico: contribución de los rasgos de personalidad, creencias de autoeficacia e intereses. *Revista Colombiana de Psicología*, *23*(1), 57-71. Recuperado de http://www.redalyc.org/articulo.oa?id=80431219003

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, New York: The Guilford Press. Recuperado de http://goo.gl/VLZzWJ

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

Engelhard Jr., G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, New York: Routledge. http://dx.doi.org/10.1007/S11336-013-9398-1

García, M. I. B., Tello, F. P. H., Abad, E. V., & Moscoso, S. C. (2007). Actitudes, hábitos de estudio y rendimiento en Matemáticas: diferencias por género. *Psicothema*, *19*(3), 413-421. Recuperado de http://goo.gl/hnqMUP

Hammouri, H. & Sabah, S. A. (2010). Analysis and assessment of the Jordan National Test for Controlling the Quality of Science Instruction (NTCQSI): A Rasch measurement perspective. *Educational Research and Evaluation*, *16*(6), 451-470. http://dx.doi.org/10.1080/09243453.2010.550469

Kleinman, M. & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling, 58*, 79-98. Recuperado de https://goo.gl/0bqiJB

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106. http://dx.doi.org/10.1.1.424.2811

Linacre, J. M. (2016). Winsteps® (Version 3.92.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Recuperado el 1 de enero de 2016 de http://www.winsteps.com/

Long, C., Wendt, H., & Dunne, T. (2011). Applying Rasch measurement in mathematics education research: Steps towards a triangulated investigation into proficiency in the multiplicative conceptual field. *Educational Research and Evaluation, 17*(5), 387-407. http://dx.doi.org/10.1080/13803611.2011.632661

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6*(1), 97-103. http://dx.doi.org/10.1007/BF01908590

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3ª ed., pp 13-103) New York, New York: MacMillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. http://dx.doi.org/10.3102/0013189X023002013

Olea, J., Abad, F. J., Ponsoda, V., & Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y comprobaciones psicométricas. *Psicothema, 16*(3), 519-525. Recuperado de http://goo.gl/8xcmfG

Organisation for Economic Cooperation and Development, OECD. (2006). The Programme for International Student Assessment (PISA). Washington, District of Columbia: NCES.

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, España: Pirámide. Recuperado de http://goo.gl/MI2dLi

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements. En M. Blegvad (Ed.), *The Danish year-book of philosophy* (pp. 58-94). Copenhagen, Denmark: Munksgaard.

Schulz, W. & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation, 17*(6), 447-464. http://dx.doi.org/ 10.1080/13803611.2011.630559

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. En K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Parks, California: Sage.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 3-24. http://dx.doi.org/ 10.1080/10705519609540026

Wright, B. D. & Stone, M. H. (2003). Five steps to science: Observing, scoring, measuring, analyzing, and applying. *Rasch Measurement Transactions*, *17*(1), 912-913. Recuperado de https://goo.gl/ZJagiH

Zalazar-Jaime, M. F., Cupani, M., & De Mier, V. (2015). Evaluation of the performance model of social cognitive theory of career: Contributions of differential learning experiences. *Bordón. Revista de Pedagogía*, *67*(4), 153-168. http://dx.doi.org/10.13042/Bordon.2015.67410

Fecha de recepción: 31 de mayo de 2016
Fecha de aceptación: 5 de septiembre de 2016