Calibración de un banco de ítems mediante el modelo de Rasch para medir razonamiento numérico, verbal y espacial*

Calibration of an Item Bank Using the Rasch Model to Measure Numeric, Verbal and Spatial Reasoning

Calibração de um banco de itens através do modelo de Rasch para medir razoamento numérico, verbal e especial

Fernanda B. Ghio[*,**]
Valeria E. Morán[**]
Sebastian J. Garrido[**]
Ana E. Azpilicueta[**]
Franco Córtez[**]
Marcos Cupani[**]

Doi: http://dx.doi.org/10.12804/revistas.urosario.edu.co/apl/a.7760

Resumen

En el ámbito educativo las pruebas de inteligencia son consideradas una de las mejores predictoras del rendimiento académico de los estudiantes. El propósito de este estudio es la adaptación de un grupo de reactivos y la conformación de un banco de ítems (BI) que permita evaluar de manera precisa y objetiva algunas aptitudes cognitivas específicas (razonamiento verbal, numérico y espacial) y generar un indicador general de inteligencia. Para ello se seleccionaron, tradujeron y administraron 255 preguntas del BI propuesto por los autores Russell y Carter (2015). La muestra estuvo

* Dirigir correspondencia a Fernanda B. Ghio. Correo electrónico: fernandabghio@gmail.com

** Instituto de Investigaciones Psicológicas, IIPSI, Unidad Ejecutora CONICET, Facultad de Psicología, Universidad Nacional de Córdoba, Argentina.

Este trabajo ha recibido apoyo financiero del Fondo para la Investigación Científica y Tecnológica (FONCYT). Préstamo BIT PICT 2016 N° 4381 y del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

Para citar este artículo: Ghio, F. B., Morán, V. E., Garrido, S. J., Azpilicueta, A. E., Córtez, F., & Cupani, M. (2020). Calibración de un banco de ítems mediante el modelo de Rasch para medir razonamiento numérico, verbal y espacial. *Avances en Psicología Latinoamericana*, *38*(1), 1-15. Doi: http://dx.doi.org/10.12804/revistas.urosario.edu.co/apl/a.7760

compuesta por 1140 estudiantes pertenecientes a la Universidad Nacional de Córdoba (Argentina), 616 del sexo femenino, 392 del sexo masculino y 132 que no reportaron el sexo, con edades comprendidas entre los 17 y 49 años (M = 20.29; DE = 3.25). Los datos se analizaron mediante el modelo de Rasch. Los resultados expresan que los ítems, en general, poseen adecuadas propiedades psicométricas tanto para las aptitudes específicas, como para los ítems que conforman un indicador general de inteligencia. Se recomienda la utilización del modelo de Rasch para la construcción o adaptación de pruebas y se discuten las implicancias de la utilización de este modelo en las pruebas de aptitudes cognitivas.

Palabras clave: rendimiento académico, inteligencia, aptitudes cognitivas específicas, modelo de Rasch, banco de ítems.

Abstract

In the educational field, intelligence tests are considered one of the best predictors of the academic performance of students. The purpose of the present study is the adaptation of an item pool and the development of an item bank (IB) that allows to accurately and objectively assess some specific cognitive abilities (verbal, numerical and spatial reasoning) and an indicator of general intelligence. For this purpose, the authors selected, translated, and administered 255 items from the Russell and Carter (2015) IB. The sample consisted of 1140 students from the National University of Córdoba (Argentina), 616 female, 392 male, and 132 who did not report the sex, with ages ranging between 17 and 49 years (M = 20.29; SD = 3.25). The results express that the items, in general, have adequate psychometric properties both for the specific skills and for general intelligence. The authors used the Rasch model for data analysis, recommended for the construction or adaptation of tests. The article discusses the implications of the application of this model in cognitive skills tests.

Keywords: Academic performance, intelligence, specific cognitive abilities, rasch model, item bank.

Resumo

No âmbito educativo as provas de inteligência são consideradas uma das melhores preditoras do rendimento acadêmico dos estudantes. O propósito deste estudo é a adaptação de um pool de reativos e a conformação de um banco de itens (BI) que permita avaliar de forma precisa e objetiva algumas competências cognitivas específicas (razoamento verbal, numérico e espacial) e gerar um indicador geral de inteligência. Para isto se selecionaram, traduziram e administraram 255 perguntas do BI proposto pelos autores Russell e Carter (2015). A amostra esteve composta por 1140 estudantes pertencentes à *Universidad Nacional de Córdoba* (Argentina), 616 do sexo feminino, 392 do sexo masculino e 132 que não reportaram o sexo, com idades entre os 17 e 49 anos (M = 20.29; DE = 3.25). Os dados se analisaram através do Modelo de Rasch. Os resultados expressam que os itens, em geral, possuem adequadas propriedades psicométricas tanto para as competências específicas quanto para os itens que conformam um indicador geral de inteligência. Recomenda-se a utilização do modelo de Rasch para a construção ou adaptação de provas e discutem-se as implicâncias da utilização deste modelo nas provas de competências cognitivas.

Palavras-chave: desempenho escolar, inteligência, competências cognitivas específicas, modelo de rasch, banco de itens.

El rendimiento académico (RA) es utilizado como un indicador del conocimiento adquirido por los estudiantes (Mingorance, Trujillo, Cáceres, & Torres, 2017). De allí, que en el ámbito educativo se utiliza para evaluar la calidad de la educación en una institución académica (Pluut, Curşeu, & Ilies, 2015; Santhya, Zavier, & Jejeebhoy, 2015). Cabe mencionar que existe un amplio bagaje de estudios que investigan la influencia de factores cognitivos (inteligencia) y no cognitivos en el RA como rasgos de personalidad, entre otras (Ackerman & Beier, 2006; Rimfeld, Kovas, Dale, & Plomin,

2016). Sin embargo, a partir de la década del 30, dichos estudios le han otorgado mayor importancia a la inteligencia como un factor que juega un rol esencial en el éxito académico (Almeida, Guisande, Primi, & Lemos, 2008; Alves, 2015; Kaya, Juntune, & Stough, 2015; Soares, Lemos, Primi, & Almeida, 2015).

Al estudiar la inteligencia, principalmente relacionada con el desempeño académico, numerosas investigaciones abordan los dominios específicos de razonamiento numérico (RN), verbal (RV) y espacial (RE) (Bresgi, Alexander, & Seabi, 2017; Schillinger, Vogel, Diedrich, & Grabner, 2018; Thomas, Rammsayer, Schweizer, & Troche, 2015; van de Weijer-Bergsma, Kroesbergen, & van Luit, 2015). El RN permite, fundamentalmente, comprender relaciones y conceptos numéricos; el RV remite a las capacidades de las personas en la comprensión de lenguaje (Gambari, Kutigi, & Fagbemi, 2014), y el RE a encontrar relaciones entre formas y figuras (Wai, Lubinski, & Benbow, 2009). Existe evidencia que estas aptitudes cognitivas específicas contribuyen a la predicción del rendimiento en dominios particulares (Barca-Enriquez et al., 2015; Solano Luengo, 2015; Pérez, Cupani, & Ayllón, 2005; Spinath, Freudenthaler, & Neubauer, 2010). Además, otros estudios demuestran que evaluando estos tres tipos de aptitudes se obtiene un indicador general de inteligencia (Lubinski, 2004).

Un aspecto que suele diferenciarse en los estudios de inteligencia y desempeño académico refiere a las puntuaciones obtenidas en cada tipo de razonamiento —aptitudes cognitivas— entre hombres y mujeres. Por un lado, existen estudios que postulan que no habría diferencias (Halpern, Beninger, & Straight, 2011; Sternberg, 2014). Por otro, hay estudios que muestran que, mientras los hombres suelen tener, en general, un mejor rendimiento en la mayoría de las medidas de inteligencia (Spinath et al., 2010), las mujeres suelen destacarse en las pruebas de dominio verbal (Kohút, Halama, Dockal, & Zitný, 2016). En este sentido, Ackerman (2006) encontró que estudios empíricos demostraron que dichas diferencias dependerán de la operacionalización del contenido con la que se construyó el instrumento de medición. Es decir, la diferencia entre los grupos dependerá del tipo de contenido o subtipo de aptitud cognitiva que esté evaluando el instrumento.

Actualmente, las principales pruebas de inteligencia como el Woodcock–Johnson Tests of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001) y la Wechsler Intelligence Scale (Wechsler, 2008, 2014) utilizan como modo de interpretación de sus puntuaciones el modelo teórico de tres estratos (Carroll, 1993). Estos instrumentos fueron diseñados desde la teoría clásica de los test (TCT) y luego calibrados mediante la teoría de respuesta al ítem (TRI). Esta última permite obtener medidas invariables, independientes de los instrumentos utilizados y de los individuos evaluados (Engelhard, 2013). Además, posibilita el análisis del funcionamiento diferencial de los ítems (DIF, por sus siglas en inglés) y la generación de pruebas adaptativas computarizadas (CATs, por sus siglas en inglés) (Reise & Waller, 2009).

En la última década, las CATs adquirieron popularidad en la medición ya que ofrecen ventajas respecto a los formatos fijos de pruebas tradicionales. Estas permiten que un instrumento se adapte al examinado en el momento en que está respondiendo. Es decir, este tipo de pruebas actualiza continuamente las estimaciones de la posición (habilidad) en el constructo de interés (rasgo latente). Dicha situación se genera a partir de las respuestas dadas por el examinado a las preguntas propuestas (van der Linden & Glas, 2000). De esta forma, si el participante responde incorrectamente a un ítem, el siguiente tendrá un nivel de dificultad menor; si responde correctamente, se le presentará un ítem de mayor dificultad (Atorresi, Lozzia, Abal, Galibert, & Aguerri, 2009). Varios estudios han demostrado que las CATs pueden limitar, en gran medida, los elementos administrados

y en efecto reducir la carga en los participantes sin pérdida de precisión de la medición (Gershon, 2005; Thompson, 2011). Además, la aplicación de este tipo de pruebas aumenta la motivación de los examinados debido a que los ítems se seleccionan considerando sus niveles de habilidad (Gibbons et al., 2008).

Un prerrequisito necesario para la construcción de una CATs es la creación de un banco de ítem (BI) que mantenga la seguridad de la prueba y, por lo tanto, su validez (Mills & Steffen, 2000; Stocking & Lewis, 2000). Un BI de alta calidad es crucial para el establecimiento de una CATs exitosa. Su generación correctamente calibrada podría posibilitar el desarrollo de nuevas pruebas lápiz y papel de longitud fija para medir la inteligencia. Los investigadores podrían seleccionar conjuntos determinados de ítems de un BI para diagnósticos específicos, por ejemplo, para la identificación de estudiantes talentosos. Además del desarrollo de pruebas paralelas, que permitan mediciones repetidas (Lozzia et al., 2015).

Considerando una de las ventajas que ofrecen los modelos basados en TRI: la posibilidad de conformar un BI con parámetros definidos (Atorresi et al., 2009), el objetivo de este trabajo fue adaptar y calibrar un conjunto de ítems para conformar un BI destinado a evaluar algunas aptitudes cognitivas específicas y generar un indicador general de inteligencia para pruebas a medida. Para lograr este fin, se seleccionaron, tradujeron y adaptaron un grupo inicial de ítems que miden RV, RN y RE propuestos por Russell y Carter (2015). Luego se procedió a la calibración de los ítems mediante el modelo de Rasch, el cual aporta importantes avances en el desarrollo de evaluaciones a gran escala, utilizadas en el ámbito educativo (Wendt, Bos, & Goy, 2011), uno de los cuales permite ubicar en una misma escala los parámetros de los ítems y las personas y, en consecuencia, obtener evaluaciones invariantes adaptadas al nivel de competencia de las personas (Prieto & Delgado, 2003).

Método

Participantes

Se realizó un muestreo no probabilístico de tipo accidental, donde participaron 1140 estudiantes, 616 de sexo femenino (54 %), 392 de sexo masculino (34.4 %) y 132 que no reportaron su sexo (11.6%), con edades comprendidas entre los 17 y 49 años (M = 20.29; DE = 3.25), que estaban estudiando en diferentes facultades de la Universidad Nacional de Córdoba (UNC): Psicología (42.5 %), Ciencias Agropecuarias (9.7 %), Medicina (1.9 %), Ciencias Económicas (30.6 %), Ciencias de la Comunicación (5 %), Arquitectura, Urbanismo y Diseño (8.2%); un 2 % no documentaron a cuál facultad pertenecían. En cuanto al año cursado, 450 participantes cursaban el Ciclo de Nivelación (39.5 %); 78, primer año (6.8 %); 363, segundo año (31.8 %); 160, tercer año (14 %); 18, cuarto año (1.6 %) y un 6.1 % no reportó su año cursado.

Instrumento

El BI es el propuesto por los autores Russell y Carter (2015). Este grupo de 800 ítems se organiza en diferentes pruebas de 40 preguntas cada una; algunas son de opción múltiple (3, 4 o 5 opciones) y otras de completar. La puntuación se realiza otorgando un punto a cada respuesta correcta y se propone una escala de corrección según cuántas preguntas respondieron bien cada uno de los sujetos. Los reactivos miden RV, RN y RE. Antes de realizar la adaptación de estas preguntas se solicitaron los permisos pertinentes y se obtuvo la licencia para utilizar los ítems solo para fines de investigación.

Procedimiento

Traducción y desarrollo de los ítems. Se comenzó realizando una primera selección de 422 preguntas, para la que se utilizaron los siguientes

criterios: (a) se incluyeron solo aquellas preguntas que evaluaran alguno de los razonamientos propuestos (numérico, verbal y espacial); (b) las preguntas debían contar con tres opciones de repuestas, tener un formato que permitiese añadir opciones de respuesta o en caso de tener más de tres opciones de repuestas, debía poder adaptarse a tres sin alterar significativamente la estructura del ítem; (c) seleccionar, en caso que fuese necesario, preguntas que utilizaran imágenes suficientemente nítidaz. A continuación, se realizó la traducción y adecuación de las preguntas en términos de formato, claridad, equivalencia de términos en español-inglés y consideraciones de estilo. Durante esta parte del proceso se contó con la colaboración de dos profesionales quienes asesoraron y corrigieron los ítems traducidos según criterios de adecuación semántica y sintáctica. Al finalizar este proceso quedaron 255 preguntas (85 para medir RN, 100 para RV y 70 para RE). En la figura 1 se presenta un ejemplo de los ítems de cada una de las capacidades evaluadas.

Diseño, montaje y producción de la prueba. Se organizaron las preguntas en diferentes cuadernillos según el contenido que apuntaba medir cada ítem (RN, RV y RE). Para facilitar la administración y calibración de todos los ítems se crearon diferentes formas de los cuadernillos de preguntas. Cada uno estaba compuesto por 25 preguntas en total (10 comunes de anclaje y 15 libres). Las opciones de respuestas correctas, variaron de ubicación de manera aleatoria en cada una de las formas. No obstante, los ítems anclas estuvieron siempre ubicados en el mismo orden en cada uno de los cuadernillos. Por otro lado, se conformó un protocolo de repuesta con espacios determinados para la elección de las mismas (A, B o C). De acuerdo con el conjunto total de preguntas que se obtuvo por cada razonamiento, quedaron establecidas cinco formas para el RN (A, B, C, D y E); seis para el RV (A, B, C, D, E y F) y cuatro para RE (A, B, C y D). En la tabla 1, se presentan cada una de las formas administradas y la cantidad de participantes que respondieron.

Administración de la prueba. Previo a la administración de los instrumentos, se solicitaron permisos a las autoridades correspondientes. Se contó con aulas con capacidades para administrar 100 a 200 protocolos en una sola sesión. Estas administraciones se realizaron en horarios regulares de clases y con la presencia del profesor a cargo. Los participantes fueron informados y dieron su consentimiento acerca del propósito del estudio, la voluntariedad de su participación y la naturaleza confidencial de los datos. Cada participante respondió tres protocolos, uno de cada capacidad cognitiva y distribuida en las distintas formas. Por ejemplo, de los 223 estudiantes que respondieron la forma E de razonamiento numérico (ver tabla 1), 28 respondieron de razonamiento verbal, la forma A; 39 la forma B; 44 la forma C; 21 la forma D; 45 la forma E, y 46 la forma F. Por último, cabe mencionar que los procedimientos del estudio, incluyendo los aspectos éticos, fueron aprobados por la Comisión Evaluadora de Proyectos de la Secretaria de Ciencia y Tecnología de la UNC.

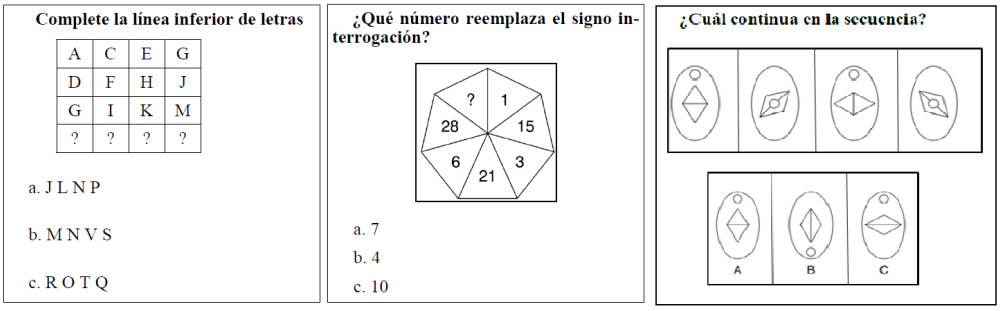

Figura 1. Ejemplo de ítems para evaluar el razonamiento verbal, numérico y espacial.

Tabla 1.
Cantidad de participantes que respondieron a cada forma de los subtest de razonamiento

Formas	Subtest de razonamiento		
	RN (n)	RV (n)	RE (n)
A	288	219	259
B	215	189	291
C	224	179	270
D	231	159	290
E	223	161	
F		230	
∑ n	1121	1137	1110

Nota: RN= Razonamiento Numérico; RV= Razonamiento Verbal; RE= Relaciones Espaciales; ∑ *n*= sumatoria de los participantes que respondieron en cada subtest de razonamiento

Análisis de datos

Los análisis se realizaron para todas las formas de RN, RV y RE, y para los ítems anclas. En un primer momento se evaluó el supuesto de unidimensionalidad mediante el Normal Ogive Harmonic Analysis Robust Method, utilizando el programa NOHARM versión 4.0. El *software* provee el Índice Tanaka (1993) de Bondad de Ajuste (GFI) como una medida de ajuste. McDonald (1989) sugiere que un puntaje de .90 es un valor aceptable y que un índice de .95 indica un buen ajuste. Con el objetivo de chequear el ajuste entre los datos y el modelo, seguimos diferentes procedimientos. Mediante el *software* WINSTEPS se calcularon dos medidas de *infit* (Mnsq y Zstd) y dos medidas de *outfit* (Mnsq y Zstd). De acuerdo a los criterios propuestos por Bond y Fox (2003) un Zstd con ajuste aceptable debe fluctuar entre valores iguales o mayores a +2 e iguales o menores a -2. Por otro lado, también se obtuvieron los índices de separación y fiabilidad de ítems y personas. El índice de separación superior a 2.0 se considera adecuado al igual que una fiabilidad asociada al índice de separación de .80 (Bond & Fox, 2003). Finalmente, se realizaron análisis de funcionamiento diferencial del ítem (DIF, por sus siglas en inglés) según el sexo de los participantes. Para aplicar el DIF se llevaron a cabo análisis *pair-wise* en los que el nivel de significación se fijó en $\alpha < .01$ y se tuvo en cuenta que el contraste del DIF debía ser superior a ≥ 0.5 *logit* (Linacre, 2016). Para este análisis, Winsteps utiliza la *t* de Welch (Linacre, 2016).

Con el fin de generar evidencias de validez del BI resultante, se realizó un análisis correlacional entre las tres capacidades cognitivas. Las correlaciones de Pearson se calcularon entre cada una de las formas (A, B, C, etc.) y de cada una de las capacidades evaluadas y del puntaje *logit* obtenido para cada uno de los participantes. Estas correlaciones se realizaron por separado según la variable sexo. Luego, se comparó cómo difieren las capacidades cognitivas según dicha variable y el área de estudio; se realizaron los estudios de prueba *t* de diferencias de medias y análisis de la varianza (ANOVA) de un factor. De manera adicional, se estimó el tamaño del efecto para ambos estudios. En la prueba *t* se consideró el estadístico *d* de Cohen, a partir de los criterios establecidos de pequeño (d = .10 a .20), mediano (d = .21 a .50) y grande (d = > .51) (Cohen, 1988). Para el ANOVA, se estimó el coeficiente eta cuadrado parcial (η^2p). Para su interpretación se consideraron valores pequeños ($\eta^2p = .01$), medianos, ($\eta_2p = .06$), y grandes ($\eta^2p = .14$) (Cohen, 1988).

Resultados

Ítems que miden RN

Forma A de RN. El valor del GFI de .95 indica que se cumple el supuesto de unidimensionalidad. En relación al ajuste de los ítems al modelo, se observó que los 25 presentaron un buen ajuste (*infit* y *outfit* Zstd $\leq \pm$ 2.0). La medida de dificultad (δi) de los ítems varió entre $-1.49 \leq \delta_i \leq 1.54$ ($M = 0.00$, $DE = 0.80$). Por su parte, el análisis de ajuste de las personas refleja que el 93 % de los patrones de respuesta se ajustaron al modelo (*infit* y *outfit* Zstd $\leq \pm$ 2.0). Los niveles de habilidad variaron entre $-3.46 \leq \theta \leq 1.86$ ($M = -0.77$, $DE = 1.10$). Los índices de separación (5.04) y fiabilidad (.96) de los ítems y los índices de separación (1.87) y fiabilidad (.78) de las personas fueron satisfactorios. Los resultados del análisis de DIF según el sexo permiten observar que dos ítems (10 y 15) presentaron un funcionamiento diferencial. En un mismo nivel de habilidad, las mujeres fueron menos capaces que los hombres de responder correctamente a estos ítems.

Forma B de RN. El GFI fue de .93. De los 25 ítems, 24 presentaron un ajuste adecuado al modelo. Los índices de dificultad de los ítems variaron entre $-1.11 \leq \delta_i \leq 1.08$ ($M = 0.00$, $DE = 0.58$). El 94.4% de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.34 \leq \theta \leq 1.24$ ($M = -1.13$, $DE = .91$). Los índices de separación (3.55) y fiabilidad (.93) de los ítems fueron satisfactorios. Los índices de separación (1.39) y fiabilidad (.66) de las personas fueron adecuados. El análisis de DIF según el sexo permitió observar que solo un ítem (24) presentó un funcionamiento diferencial.

Forma C de RN. El GFI fue de .89, valor próximo al punto de corte propuesto. De los 25 ítems, 24 presentaron un ajuste adecuado. El índice de dificultad de los ítems varió entre $-1.35 \leq \delta_i \leq 1.49$ ($M = 0.00$, $DE = 0.61$). El 94.5 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.35 \leq \theta \leq 1.24$ ($M = -0.89$, $DE = 1.04$). Los índices de separación (3.40) y fiabilidad (.92) de los ítems fueron satisfactorios. Los índices de separación (1.73) y fiabilidad (.75) de las personas fueron adecuados. No se observó un funcionamiento diferencial de ninguno de los 25 ítems según el sexo.

Forma D de RN. El GFI fue de .88, no superó el punto de corte propuesto. De los 25 ítems, 23 presentaron un ajuste adecuado. El índice de dificultad de los ítems varió entre $-1.73 \leq \delta_i \leq 1.04$ ($M = 0.00$, $DE = 0.59$). El 94.9 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.36 \leq \theta \leq 1.48$ (M = -1.02, DE = 0.91). Los índices de separación (2.83) y fiabilidad (.89) de los ítems y los de separación (1.44) y fiabilidad (.67) de las personas fueron adecuados. No se observó un funcionamiento diferencial de los ítems según el sexo.

Forma E de RN. El GFI fue de .91. De los 25 ítems, 23 presentaron un ajuste adecuado. El índice de dificultad varió entre $-1.35 \leq \delta_i \leq 0.80$ ($M = 0.00$, $DE = 0.58$). El 95.5 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.34 \leq \theta \leq 0.63$ ($M = -1.26$, $DE = 0.87$). Los índices de separación (3.18) y fiabilidad (.91) de los ítems y los de separación (1.23) y fiabilidad (.60) de las personas fueron adecuados. No se observó un funcionamiento diferencial de los ítems según el sexo.

Ítems que miden RV

Forma A de RV. El GFI fue de .92. Los 25 ítems presentaron un ajuste adecuado al modelo. El índice de dificultad de los ítems varió entre $-2.29 \leq \delta_i \leq 1.72$ ($M = 0.00$, $DE = 1.17$). El 96.8 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-2.99 \leq \theta \leq 2.07$ ($M = 0.16$, $DE = 1.10$). Tanto los índices de separación (6.89) y fiabilidad (.98) de los ítems como los de separación (1.98) y fiabilidad (.80) de las personas fueron satisfactorios. Los ítems 19 y 24 presentaron DIF.

Forma B de RV. El GFI fue de .90. De los 25 ítems, 24 presentaron un ajuste adecuado al modelo. El índice de dificultad de los ítems varió entre $-1.65 \leq \delta_i \leq 2.24$ ($M = 0.00$, $DE = 1.11$). El 94.2 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.63 \leq \theta \leq 3.77$ ($M = 1.14$, $DE = 0.65$). Los índices de separación (5.91) y fiabilidad (.97) de los ítems y los de separación (1.95) y fiabilidad (.79) de las personas fueron satisfactorios. Los ítems 18 y 25 presentaron DIF.

Forma C de RV. El GFI fue de .80 no fue adecuado. Los 25 ítems presentaron un ajuste adecuado al modelo y la dificultad varió entre $-2.08 \leq \delta_i \leq 2.11$ ($M = 0.00$, $DE = 1.15$). El 97.8 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.71 \leq \theta \leq 2.07$ ($M = 0.36$, $DE = 0.99$). Tanto los índices de separación (6.13) y fiabilidad (.97) de los ítems como los de separación (1.73) y fiabilidad (.75) de las personas fueron satisfactorios. No se observó DIF.

Forma D de RV. El GFI de .79 no fue adecuado. Los 25 ítems presentaron un ajuste adecuado al modelo. El índice de dificultad varió entre $-1.53 \leq \delta_i \leq 1.83$ ($M = 0.00$, $DE = 1.05$). El 94.3 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.59 \leq \theta \leq 2.40$ ($M = 0$ .31; $DE = 1.18$). Los índices de separación (5.26) y fiabilidad (.97) de los ítems, así como los índices de separación (2.12) y fiabilidad (.82) de las personas resultaron satisfactorios. El ítem 4 presento DIF.

Forma E de RV. El GFI fue de 0.79. Los 25 ítems presentaron un ajuste adecuado al modelo y los índices de dificultad variaron entre $-2.10 \leq \delta_i \leq 1.91$ ($M = 0.00$, $DE = 0.98$). El 95.7 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-2.81 \leq \theta \leq 1.64$ ($M = -0.20$, $DE = 0.92$). Tanto los índices de separación (5.11) y fiabilidad (0.96) de los ítems como los de separación (1.66) y fiabilidad (0.73) de las personas fueron adecuados. No se observó DIF.

Forma F de RV. El GFI fue 0.86. Los 25 ítems presentaron un ajuste adecuado al modelo. El índice de dificultad de los ítems varió entre $-1.66 \leq \delta_i \leq 3.02$ ($M = 0.00$, $DE = 1.28$). El 94.8 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-2.94 \leq \theta \leq 2.62$ ($M = 0.54$, $DE = 1.11$). Los índices de separación (7.35) y fiabilidad (.98) de los ítems y los de separación (1.88) y fiabilidad (.78) de las personas fueron satisfactorios. No se observó DIF.

Ítems que miden RE

Forma A de RE. El GFI fue de .94. Veintitrés ítems presentaron un ajuste adecuado. El índice de dificultad de los ítems varió entre $-2.29 \leq \delta_i \leq 1.72$ ($M = 0.00$, $DE = 1.17$). El .93.1% de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.35 \leq \theta \leq 0.82$ (M = -1.33, DE = 0.99). Los índices de separación (3.66) y fiabilidad (.93) de los ítems, así como los de separación (1.40) y fiabilidad (.66) de las personas fueron adecuados. No se observó DIF.

Forma B de RE. El GFI fue de .93. Los 25 ítems presentaron un ajuste adecuado al modelo. El índice de dificultad de los ítems varió entre $-1.08 \leq \delta_i \leq 1.05$ ($M = 0.00$, $DE = 0.53$). El 94.2 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.31 \leq \theta \leq 2.55$ ($M = -1.01$, $DE = 1.07$). Tanto los índices de separación (3.49) y fiabilidad (.92) de los ítems, como los índices de separación (1.70) y fiabilidad (.74) de las personas fueron adecuados. Los ítems 17 y 18 presentaron DIF.

Forma C de RE. El GFI fue de .93. Los 25 ítems presentaron un ajuste adecuado al modelo. El índice de dificultad de los ítems varió entre $-0.89 \leq \delta_i \leq 1.27$ ($M = 0.00$, $DE = 0.53$). El 92.6 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.31 \leq \theta \leq 0.61$ (M = -1.34, DE = 0.85). Los índices de separación (3.20) y fiabilidad (.91) de los ítems

fueron satisfactorios, mientras que los de separación (1.16) y fiabilidad (.57) de las personas no lo fueron. El ítem 3 presentó DIF.

Forma D de RE. El GFI fue de .90. De los 25 ítems, 24 presentaron un ajuste adecuado al modelo. El índice de dificultad de estos varió entre $-1.16 \leq \delta_i \leq 0.74$ ($M = 0.00$, $DE = 0.48$). El 94.1 % de los patrones de respuesta se ajustó al modelo. Los niveles de habilidad variaron entre $-3.29 \leq \theta \leq 1.45$ ($M = -1.04$, $DE = 0.95$). Los índices de separación (3.24) y fiabilidad (.91) de los ítems resultaron satisfactorios, mientras que los de separación (1.48) y fiabilidad (.69) de las personas fueron levemente adecuados. No se observó DIF.

Ítems anclas para medir un factor general

El valor del GFI de .91 indica que se cumple el supuesto de unidimensionalidad. En relación al ajuste de los ítems al modelo, se observó que de los 30 ítems, 29 de ellos presentaron un buen ajuste (*infit* y *outfit* Zstd $\leq \pm 2.0$). El ítem 27 que evalúa RE no se ajustó al modelo. La medida de dificultad (δi) de los ítems varió entre $-2.51 \leq \delta i \leq 1.43$ ($M = 0.00$; $DE = 0.97$). Por su parte, el 95.1% de los patrones de respuesta se ajustaron al modelo. Los niveles de habilidad variaron entre $-1.63 \leq \theta \leq 3.09$ ($M = -0.56$, $DE = 0.72$). Los índices de separación (13.75) y fiabilidad (.99) de los ítems fueron satisfactorios. Por otro lado, el índice de separación de las personas (1.33) y de fiabilidad (.64) no fue satisfactorio. No se observó DIF.

Evidencias de validez del BI

El BI resultante quedó conformado por 225 preguntas que se ajustaron adecuadamente a los datos. Los índices de dificultad de estos ítems variaron entre $-1.89 \leq \delta i \leq 3.02$ ($M = 0.42$; $DE = 0.82$). En la figura 2 se presentan los valores de dificultad diferenciados por capacidad cognitiva.

Se puede observar que los ítems de RV abarcan un nivel de dificultad más amplio (aproximadamente entre -2.0 a 2.0 *logit*), seguido de los de RN (-1.5 a 1.5); en último lugar se encuentran los de RE (-1.0 a 1.0). Con relación al nivel de habilidad de los participantes, los puntajes *logit* variaron entre $-5.06 \leq \theta \leq 3.77$ ($M = -0.67$, $DE = 1.33$).

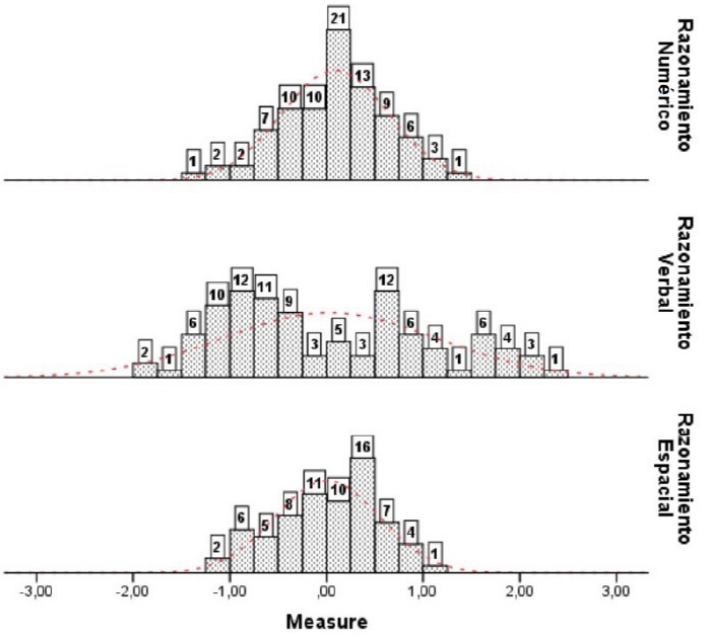

Figura 2. Gráfico de histograma que muestra la distribución de los índices de dificultad de los ítems separados por capacidad cognitiva.

Correlación entre las capacidades. En la tabla 2 se observa cómo se correlacionan cada una de las formas (A, B, C, etc.) en cada una de las capacidades evaluadas. Se puede destacar que, aunque las correlaciones observadas son positivas, las mismas varían de un tamaño del efecto pequeño a moderado y con valores más grandes en la muestra masculina.

Diferencias de grupo según sexo y área de estudio. La prueba *t* de diferencia de medias no arrojó diferencias significativas en las tres capacidades cognitivas [RN (t = -0.972, p = .33), RV (t = 0.72, p = .47), RE (t = 0.579, p = .56)] en relación al sexo de los participantes. Por otro lado, los resultados del ANOVA indicaron una asociación en RN [F (2, 1000) = 8.07, p = .00, η^2p = .02], RV [F (2, 1048) = 4.49, p = .01, η^2p = .01] y RE [F (2, 1027) = 6.99, p = .01, η^2p = .01] con las áreas de estudio. La comparación *post hoc* mediante la prueba de Bonferroni indicó que los estudiantes de ciencias sociales obtuvieron puntajes más bajos en razonamiento numérico en comparación a los estudiantes de ciencias naturales, básicas, aplicadas y de la salud, aunque entre estos grupos no se observaron diferencias. Con relación a las capacidades verbales, el grupo de estudiantes de ciencias sociales obtuvo un rendimiento superior a los estudiantes de ciencias naturales, básicas y aplicadas. No se observó diferencia entre los otros grupos. Finalmente, en relación con el razonamiento espacial, se observó que los estudiantes de ciencias de la salud obtuvieron un puntaje mayor a los estudiantes de ciencias sociales. No se observó otras diferencias entre los otros grupos.

Tabla 2.
Correlación entre los puntajes logit obtenidos en cada una de las formas de los subtest de razonamiento separadas por la variable sexo

Subtest de Razonamiento	Femenino n (616)			Masculino n (392)		
	RN	RV	RE	RN	RV	RE
RN						
Forma A	-	.03	.19*	-	.53**	.41**
Forma B	-	.18	.20*	-	.26	.52**
Forma C	-	.09	.18	-	.12	.06
Forma D	-	.06	.49**	-	.53**	.20
Forma E	-	.06	.08	-	.29*	.40**
RV						
Forma A	.05	-	.31**	.24*	-	.37**
Forma B	.11	-	.23*	.24*	-	.39**
Forma C	.06	-	.04	.40**	-	.24*
Forma D	.32*	-	.33**	.21*	-	.10
Forma E	.08	-	.14	.48**	-	.26*
Forma F	.10	-	.13	.38**	-	.34**
RE						
Forma A	.28**	.26**	-	.42**	.19	-
Forma B	.02	.37**	-	.24*	.36**	-
Forma C	.25**	.04	-	.24*	.34**	-
Forma D	.36**	.12	-	.33**	.26*	-
RN	-			-		
RV	.10*	-		.34**	-	
RE	.22**	.20**	-	.30**	.30**	-

Nota: RV= Razonamiento Verbal; RN= Razonamiento Numérico; RE= Relaciones Espaciales.

Discusión

El objetivo de este estudio fue adaptar un grupo de ítems, evaluar sus propiedades psicométricas mediante el modelo de Rasch y conformar un BI que evaluara el RN, el RV y el RE que permitiera establecer un indicador general de inteligencia. Para ello se contó con un grupo inicial de 800 ítems de los cuales se adecuaron 255 para su calibración. Poder contar con un BI calibrado es un paso previo para poder realizar CATs. Esta metodología de evaluación tiene grandes ventajas en relación con las pruebas de formatos fijos. Además, posibilita administraciones en menor tiempo y con un mayor nivel de precisión.

En líneas generales, 225 ítems de los 255 que evalúan los distintos tipos de razonamiento presentaron adecuadas propiedades psicométricas. En efecto, este conjunto se ajustó adecuadamente a los datos y no presentó un funcionamiento diferencial según el sexo de los participantes. El 98% de los reactivos tuvo un nivel de dificultad que varía aproximadamente entre ± 2 *logit*. Esto indica que para poder medir niveles más bajos ($-3.0 \leq \theta \geq -2.0$) o más alto ($3.0 \geq \theta \leq 2.0$) es necesario agregar nuevos ítems que permitan estimar estos niveles de habilidad, aunque es pertinente destacar que en esta muestra en particular solo un 1.3 % presentó valores mayores a 2 *logit*.

Un análisis en detalle por capacidad específica permite observar que solo seis ítems de RN (7.6 % de los 85 ítems iniciales), siete de RV (7.5 % de los 100 ítems iniciales) y siete ítems de RE (11.1 % de los 70 ítems iniciales) no presentaron un ajuste adecuado o presentaron DIF. Los ítems de RV permitieron abarcar un rango mayor de dificultad ($-1.89 \leq \delta i \leq 3.02$) y medir con precisión a un 96 % de los participantes de este estudio ($-5.06 \leq \theta \leq 3.77$). En RN los índices de dificultad variaron en un rango menor ($-1.40 \leq \delta i \leq 1.40$) y posibilitaron medir adecuadamente a un 62% de los participantes ($-4.71 \leq \theta \leq 1.98$). En esta última capacidad, solo cinco participantes obtuvieron valores mayores a 1.40 *logit*, mientras que un 37.6 % de los participantes obtuvieron valores menores a -1.4 *logit*. Los ítems de RE mostraron índices de dificultad que variaron entre $-1.16 \leq \delta i \leq 1.13$ y permitieron medir adecuadamente a un 49.4 % de los participantes. En este caso, el 50.6 % de los participantes obtuvieron valores menores a -0.16 *logit* y solo siete participantes presentaron valores de habilidad superior a 1.13 *logit*. Por lo tanto, sería necesario contar con un número mayor de ítems que permitan estimar los niveles bajos de habilidad con mayor precisión.

En relación con la validez del BI, en líneas generales, se pudo establecer la unidimensionalidad de los ítems tanto para cada una de las formas por separado, como para los ítems anclas. Los adecuados índices de bondad de ajuste y, en consecuencia, la influencia de un rasgo para cada tipo de razonamiento permiten afirmar que este instrumento está midiendo los constructos evaluados (RN, RV y RE). De igual forma, las correlaciones entre las tres capacidades cognitivas resultaron positivas, con un tamaño del efecto que varió de pequeño a mediano, lo que permite sostener el concepto de la existencia de un indicador general de inteligencia (Lubinski, 2004; Spearman, 1927).

Con relación al análisis según el sexo, en la muestra masculina las relaciones entre las tres capacidades (correlación promedio de .31) fue levemente mayor a la muestra femenina (correlación promedio de .17), aunque en ambos casos con un tamaño del efecto moderado. Si bien las correlaciones entre las capacidades difieren según el sexo de los participantes, no se observaron diferencias significativas entre varones y mujeres en relación con la media promedio de las habilidades. Dicho comportamiento podría ser producto de la conformación de los grupos o del contenido evaluado en las pruebas de estas capacidades cognitivas específicas (Colom, García, Juan-Espinosa & Abad, 2002).

Con relación a las áreas de estudio se observaron diferencias entre los tres grupos. Los estudiantes de carreras relacionadas a las ciencias sociales

presentan más habilidades en lo verbal, los de ciencias naturales en habilidades numéricas y los alumnos de carreras afines a ciencias de la salud en habilidades espaciales. En cuanto a los ítems que presentaron DIF en RN y RE se pudo observar que, en un mismo nivel de habilidad, la mayoría de ellos fueron más difíciles para las mujeres que para los hombres, mientras que en RV, fueron más difíciles para los hombres que para las mujeres.

Si bien los resultados obtenidos son alentadores, existen limitaciones y desafíos a mencionar en este trabajo. En primer lugar, a nivel general, el grupo inicial de ítems analizados no posibilitaría evaluar satisfactoriamente a personas con niveles bajos de habilidad por lo que se hace imprescindible contar con nuevos ítems que cubran este segmento. En segundo lugar, sería pertinente incrementar la cantidad de estudiantes pertenecientes a carreras como ingeniería, matemática o física, para poder contar con participantes con un mayor nivel de habilidad en razonamientos numéricos y espaciales que permitan una cobertura más representativa. En tercer lugar, debería realizarse un análisis de los distractores para establecer si los participantes utilizaron los distractores de forma equitativa o alguno presentó un funcionamiento inesperado (Hammouri & Sabah, 2010).

En otro orden, los resultados de este trabajo tienen implicancias tanto metodológicas como prácticas. Las primeras se refieren a la utilización del Modelo de Rasch como un método que permite detectar la capacidad de los ítems para medir diferentes niveles de habilidad y, en consecuencia, determinar el conjunto de reactivos óptimo para cubrir el continuo de una prueba. Además, proporciona un análisis detallado de los patrones de respuesta individuales que reflejan los procesos de razonamiento de los individuos involucrados. La segunda apunta a que el uso de la tecnología CATs proporciona una retroalimentación empírica inmediata sobre la calidad de la recolección de datos, sumado a las ventajas en economía de recursos y tiempo.

Además, entre las implicancias prácticas se puede mencionar que los ítems calibrados pasarán a formar parte de un BI. Dicho sistema de almacenamiento introduce flexibilidad en la evaluación ya que se puede adaptar a las necesidades de los investigadores y los sistemas educativos, sumado a que las evaluaciones pueden ser comparables en diversos contextos y aplicaciones. También, permitiría el seguimiento del conocimiento de un alumno, realizar un diagnóstico de la cantidad y calidad de contenido adquirido, especificar qué conceptos resultan más dificultosos e incorporar nuevas alternativas de aprendizaje, afianzando y optimizando los métodos educativos actuales.

Referencias

Ackerman, P. L. (2006). Cognitive sex differences and mathematics and science achievement. *American Psychologist, 61*(7), 722-723. Doi:10.1037/0003-066x.61.7.722

Ackerman, P. L., & Beier, M. E. (2006). Determinants of domain knowledge and independent study learning in an adult sample. *Journal of Educational Psychology, 98*(2), 366 -381. Doi: 10.1037/0022-0663.98.2.366

Almeida, L. S., Guisande, M. A., Primi, R., & Lemos, G. (2008). Contribuciones del factor general y de los factores específicos en la relación entre inteligencia y rendimiento escolar. *European Journal of Education and Psychology, 1*(3), 5-16. Doi: 10.30552/ejep.v1i3.13

Alves, A. F. (2015). Inteligência e rendimento escolar: Implicações para a sala de aula. *Revista de Estudios e Investigación en Psicología y Educación, 2*(2), 113-121. Doi: 10.17979/reipe.2015.2.2.1329

Atorresi, I., Lozzia, G., Abal, F., Galibert, S., & Aguerri, M. (2009). Teoría de respuesta al ítem: conceptos básicos y aplicaciones para la medición de constructos psicológicos. *Revista Argentina de Clínica Psicológica*, *18*(2), 179-

188. Recuperado de https://www.redalyc.org/pdf/2819/281921792007.pdf

Barca-Enriquez, E., Brenlla, J. C., Peralbo, M., Almeida, L. S., Porto, A., & Barca, A. (2015). Habilidades cognitivas, autoeficacia y estrategias de aprendizaje: indicadores y determinantes del rendimiento académico en el alumnado de educación secundaria. R*evista de Estudios e Investigación en Psicología y Educación,* (1), 083-089. Doi: 10.17979/reipe.2015.0.01.460

Bresgi, L., Alexander, D. L. M., & Seabi, J. (2017). The predictive relationships between working memory skills within the spatial and verbal domains and mathematical performance of grade 2 South African learners. *International Journal of Educational Research, 81*, 1-10. Doi: 10.1016/j.ijer.2016.10.004

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. Doi: 10.1111/j.1745-3984.2003.tb01103.x

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambrigde: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge.

Colom, R., García, L. F., Juan-Espinosa, M., & Abad, F. J. (2002). Null sex differences in general intelligence: Evidence from the WAIS-III. *Spanish Journal of Psychology*, *5*(1), 29-35. Doi: 10.1017/s1138741600005801

Engelhard, G. Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Gambari, A. I., Kutigi, A. U., & Fagbemi, P. O. (2014). Effectiveness of computer-assisted pronunciation teaching and verbal ability on the achievement of senior secondary school students in oral English. *Gist Education and Learning Research Journal, 8*, 11-28. Recuperado de https://gistjournal.unica.edu.co/index.php/gist/article/view/111

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109-127.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., … Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*(4), 361-368. Doi: 10.1176/appi.ps.59.4.361

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

Halpern, D. F., Beninger, A. S., & Straight, C. A. (2011). Sex differences in intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *Cambridge handbooks in psychology. The Cambridge handbook of intelligence* (pp. 253-272). Cambridge University Press. Doi: 10.1017/CBO9780511977244.014

Kaya, F., Juntune, J., & Stough, L. (2015). Intelligence and its relationship to achievement. *Elementary Education Online*, *14*(3), 1060-1078. Doi: 10.17051/io.2015.25436

Kohút, M., Halama, P., Dockal, V., & Zitný, P. (2016). Gender differential item functioning in Slovak version of intelligence structure test 2000-revised. *Studia Psychologica, 58*(3), 238-250. Doi: 10.21909/sp.2016.03.720

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

van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic Publishers. Doi: 10.1007/0-306-47531-6

Lozzia, G. S., Abal, F. J. P., Blum, G. D., Aguerri, M. E., Galibert, M. S., & Atorresi, H. F. (2015). Construcción de un banco de ítems de analogías verbales como base para un test adaptativo informatizado. *Revista Mexicana de Psicología*, *32*(2), 134-148. Recuperado de https://www.redalyc.org/pdf/2430/243045364004.pdf

Lubinski, D. (2004). Introduction to the Special Section on Cognitive Abilities: 100 Years After Spearman's (1904) "'General Intelligence,' Objectively Determined and Measured". *Journal of Personality and Social Psychology*, *86*(1), 96-111. Doi: 10.1037/0022-3514.86.1.96

Solano Luengo, L., O. (2015). *Rendimiento académico de los estudiantes de secundaria obligatoria y su relación con las aptitudes mentales y las actitudes ante el estudio*. (Tesis doctoral, Universidad Nacional de Educación a Distancia, España). Recuperado de http://e-spacio.uned.es/fez/view/tesisuned:Educacion-Losolano

McDonald, R. P. (1989). An index of goodness-of-fit based on non-centrality. *Journal of Classification, 6*(1), 97-103. Doi:10.1007/bf01908590

Mills C.N., Steffen M. (2000). The GRE computer adaptive test: Operational issues. En W. J. van der Linden & C. A. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75-99). Springer. Doi:10.1007/0-306-47531-6_4

Mingorance, A. C., Trujillo, J. M., Cáceres, P., & Torres, C. (2017). Mejora del rendimiento académico a través de la metodología de aula invertida centrada en el aprendizaje activo del estudiante universitario deficiencias de la educación. *Journal of Sport and Health Research, 9*(1), 129-136. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=6026403

Pérez, E., Cupani, M., & Ayllón, S. (2005). Predictores de rendimiento académico en la escuela media: habilidades, autoeficacia y rasgos de personalidad. *Avaliação Psicológica, 4*(1), 1-11. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=6674814

Pluut, H., Curşeu, P. L., & Ilies, R. (2015). Social and study related stressors and resources among university entrants: Effects on well-being and academic performance. *Learning and Individual Differences, 37*, 262-268. Doi: 10.1016/j.lindif.2014.11.018

Prieto, G., & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, *15*(1), 94-100. Recuperado de http://www.psicothema.com/pdf/1029.pdf

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48. Doi: 10.1146/annurev.clinpsy.032408.153553

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of Personality and Social Psychology, 111*(5), 780-789. Doi: 10.1037/pspp0000089

Russell, K., & Carter, P. (2015). *Ultimate IQ tests: 1000 practice test questions to boost your brainpower*. Londres: Kogan Page Publishers.

Santhya, K. G., Zavier, A. J. F., & Jejeebhoy, S. J. (2015). School quality and its association with agency and academic achievements in girls and boys in secondary schools: Evidence from Bihar, India. *International Journal of Educational Development, 41*, 35-46. Doi: 10.1016/j.ijedudev.2014.12.002

Schillinger, F. L., Vogel, S. E., Diedrich, J., & Grabner, R. H. (2018). Math anxiety, intelligence, and performance in mathematics: Insights from the German adaptation of the Abbreviated Math Anxiety Scale (AMAS-G). *Learning and Individual Differences, 61*, 109-119. Doi: 10.1016/j.lindif.2017.11.014

Soares, D. L., Lemos, G. C., Primi, R., & Almeida, L. S. (2015). The relationship between intelligence and academic achievement throughout middle school: The role of students' prior academic performance. *Learning and Individual Differences, 41*, 73-78. Doi: 10.1016/j.lindif.2015.02.005

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. Londres: Macmillan.

Spinath, B., Freudenthaler, H. H., & Neubauer, A. C. (2010). Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation. *Personality and Individual Differences*, *48*(4), 481-486. doi: 10.1016/j.paid.2009.11.028

Sternberg, R.J. (2014). Teaching about the nature of intelligence. *Intelligence*, *42*, 176-179. Doi: 10.1016/j.intell.2013.08.010

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. En: W. J. van der Linden & C. A. Glas (Eds.), *Computerized adaptive testing: theory and practice* (163-182). Springer. Doi:10.1007/0-306-47531-6_9

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

Thomas, P., Rammsayer, T., Schweizer, K., & Troche, S. (2015). Elucidating the functional relationship between working memory capacity and psychometric intelligence: A fixed-links modeling approach for experimental repeated-measures designs. *Advances in Cognitive Psychology, 11*(1), 3-13. Doi: 10.5709/acp-0166-6

Thompson, N. (2011). *Advantages of computerized adaptative testing. Assessment systems*. Recuperado de http://www.assess.com/wp-content/uploads/2014/03/Advantages-of-CAT-Testing.pdf

van de Weijer-Bergsma, E., Kroesbergen, E. H., & van Luit, J. E. (2015). Verbal and visual-spatial working memory and mathematical ability in different domains throughout primary school. *Memory & Cognition, 43*(3), 367-378. Doi: 10.3758/s13421-014-0480-4

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817-835. Doi: 10.1037/a0016127

Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation*, *17*(6), 419-446. Doi: 10.1080/13803611.2011.634582

Wechsler, D. (2008). *Wechsler adult intelligence scale-fourth edition (WAIS-IV*). San Antonio, TX: Harcourt Assessment

Wechsler, D. (2014). *WISC-V: Administration and scoring manual*. Bloomington: NCS Pearson Incorporated.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

Recibido: mayo 30, 2019
Aprobado: octubre 16, 2019