Prueba para evaluar conocimiento en Leyes: análisis de los ítems mediante la aplicación del modelo de Rasch

Fernanda Belén Ghio
Instituto de Investigaciones Psicológicas, IIPsi, Unidad Ejecutora CONICET Facultad de Psicología, Universidad Nacional de Córdoba, Enfermera Gordillo s/n, Córdoba Capital, Argentina

Ana Estefanía Azpilicueta
Instituto de Investigaciones Psicológicas, IIPsi, Unidad Ejecutora CONICET Facultad de Psicología, Universidad Nacional de Córdoba, Enfermera Gordillo s/n, Córdoba Capital, Argentina

Marcos Cupani
Instituto de Investigaciones Psicológicas, IIPsi, Unidad Ejecutora CONICET Facultad de Psicología, Universidad Nacional de Córdoba, Enfermera Gordillo s/n, Córdoba Capital, Argentina

Valeria Estefanía Morán
Instituto de Investigaciones Psicológicas, IIPsi, Unidad Ejecutora CONICET Facultad de Psicología, Universidad Nacional de Córdoba, Enfermera Gordillo s/n, Córdoba Capital, Argentina

Sebastian Jesús Garrido
Instituto de Investigaciones Psicológicas, IIPsi, Unidad Ejecutora CONICET Facultad de Psicología, Universidad Nacional de Córdoba, Enfermera Gordillo s/n, Córdoba Capital, Argentina

Resumen

La evaluación del rendimiento académico de los estudiantes universitarios resulta un elemento fundamental para medir la calidad educativa en la enseñanza superior. La forma más directa de obtener dicha calificación es a través de exámenes o pruebas de medición, sin embargo si consideramos que son los docentes los que determinan la forma de evaluación, dichos instrumentos pueden presentar falencias en su elaboración. Por lo cual, en la Facultad de Psicología de la Universidad Nacional de Córdoba se está construyendo un Test que pretende estimar el rendimiento académico de los estudiantes universitarios. Razón por la cual en este trabajo se aplica el modelo de Rasch para evaluar las propiedades psicométricas de 31 ítems del Nivel I del dominio de conocimiento en Leyes. La muestra estuvo compuesta por 170 estudiantes con edades comprendidas entre 19 y 60 años (M = 24,59; DE = 6,22). En líneas generales, los ítems presentan buenas propiedades psicométricas. Este trabajo realiza aportes significativos en el ámbito de la medición y evaluación en nuestro medio, viendo la necesidad de evaluar el amplio rango de conocimientos y habilidades que deben adquirir los estudiantes en el ámbito universitario.

Palabras clave: test de conocimiento general- construcción de test- banco de ítems- teoría de respuesta al ítem- leyes

Abstract

The evaluation of the academic performance of university students is a fundamental element to measure the quality of education in a higher education. The most direct way to obtain this qualification is through exams or probing tests, however if we consider that teachers are the ones who determine the form of evaluation, these instruments may present shortcomings in their elaboration. Therefore, in the School of Psychology of the National University of Cordoba is being built a test that attempts to estimate the academic performance of university students. Reason for which in this work the Rasch model is applied to evaluate the psychometric properties of 31 items of Level I of the knowledge domain in Laws. The sample consisted of 170 students aged between 19 and 60 years (M = 24.59, DE = 6.22). In general, the items have good psychometric properties. This work makes significant contributions in the field of measurement and evaluation in our environment, based on the need to evaluate the wide range of knowledge and skills that students must achieve in the university environment.

Keywords: General Knowledge Testing- Test Construction- Item bank- Item Response Theory- Laws

Introducción

La evaluación del conocimiento adquirido por los estudiantes es de gran importancia en el proceso de aprendizaje (Shavelson, Zlatkin-Troitschanskaia, & Mariño, 2018). Comúnmente, esta evaluación se realiza a través de exámenes que permiten estimar el rendimiento académico (RA). En principio, los exámenes resultan útiles para supervisar el RA de los alumnos y para monitorear si una institución alcanza los objetivos curriculares establecidos (Vargas, 2007; Arancibia, 1997). Sin embargo, estas pruebas suelen tener limitaciones debido a que su elaboración está a cargo de los profesores, quienes proponen sus propios criterios y procedimientos de calificación (Navas, Sampascual, & Santed, 2003).

Para contrarrestar las limitaciones de los exámenes tradicionales surgieron pruebas estandarizadas que permitieron valorar el nivel de aprendizaje de un gran número de estudiantes (Martínez Rizo, 2009). Estas pruebas encuentran sustento en los denominados Bancos de ítems (BI), definidos como un conjunto de ítems que permiten valorar un rasgo o habilidad. El almacenamiento de los reactivos en dichos bancos permite que el evaluador elija el o los ítems que mejor se adapten a los objetivos de la evaluación (Olea, Ponsoda & Prieto, 1999).

Cabe mencionar que a nivel internacional existen diversas evaluaciones a gran escala basadas en bancos de ítems. El estudio PISA es un claro ejemplo de este tipo de evaluaciones, dicho estudio se realiza cada tres años con el objetivo de ofrecer información del rendimiento educativo de los estudiantes de nivel secundario y realizar análisis comparativos internacionales. De igual forma, en distintos países existen pruebas que permiten comparar resultados y evaluar los niveles de aprendizaje dentro de su sistema educativo, por ejemplo en EE.UU está la prueba NAEP (National Assessment of Educational Progress, Evaluación Nacional de Progreso Educativo, en español) (Martínez Arias, 2006), en México existe la prueba ENLACE (Examen Nacional de Logro Académico en Centros Escolares) y la EXCALE (Examen de Calidad y Logro Educativo) (Martínez Rizo, 2015).

Si bien en la República Argentina en el nivel secundario se cuenta con un Sistema Nacional de Evaluación de la Calidad Educativa, en el nivel universitario no se tiene un sistema estandarizado de evaluación académica. Por este motivo, para evaluar el conocimiento de los estudiantes universitarios de la ciudad de Córdoba (Argentina), se está construyendo un banco de ítems para el Test de Conocimiento General (TCG) compuesto por 20 dominios de conocimiento. Hasta el momento, los ítems e instrumentos construidos para evaluar los dominios de Biología, Historia, Economía, Literatura, Leyes y Psicología del TCG se analizaron a partir de Teoría Clásica de los Test , con resultados psicométricos satisfactorios (Cupani et al., 2016).

Aunque se utiliza Teoría Clásica de los Test para la construcción de BI, hace algunos años se recomienda utilizar modelos teóricos basados en TRI, entre ellos el modelo de Rasch. Dicha sugerencia surge no sólo para suplir ciertas falencias presentes en la TCT (Iraurgi, Lozano,

González-Saiz, & Trujols, 2008), sino porque aplicando modelos de TRI podremos obtener estimaciones de los parámetros de los ítems independientes de la muestra de sujetos que utilizamos para calibrarlos (Olea, Ponsoda, Prieto, 1999). Por lo cual, en este trabajo se aplico el Modelo de Rasch (TRI) a los ítems del Test de Conocimiento en Leyes (Nivel I) para realizar, por un lado, el ajuste de los datos al modelo evaluando dos de los supuestos fundamentales de la TRI (independencia local y objetividad específica) y por el otro, realizando la calibración de los ítems, estimando los parámetros de los ítems y precisión de las estimaciones.

Método

Muestra

El test se administró a 170 estudiantes de la carrera de abogacía, 102 mujeres (60%) y 68 varones (40%), con edades comprendidas entre los 19 y 60 años (M = 24,59; DE = 6,22). Respondieron al instrumento alumnos de la carrera de abogacía de la Facultad de Derecho de la Universidad Nacional de Córdoba (28,8%), de la Universidad Blas Pascal (63,5%) y de la Universidad Siglo 21 (7,6%) de la ciudad de Córdoba, Argentina.

Instrumento

Del pool de preguntas construidas para la conformación del BI del TCG, se seleccionaron para este estudio 31 ítems del Nivel I del dominio de Leyes (Cupani et al., 2016). El instrumento se conformó considerando que los ítems se ubican por nivel de dificultad ascendente, es decir comenzando con los ítems con un nivel de dificultad baja, luego los de dificultad moderada y por último los ítems más difíciles. A su vez se estableció que la respuesta correcta se ubicó aleatoriamente. Conjuntamente con el cuadernillo de preguntas se creó un protocolo de respuesta para que computen la respuesta seleccionada y completaran sus datos sociodemográficos (sexo, edad, materias y año de cursado). El instrumento reportó un Kuder Richardson 20 (KR-20) de .86.

Procedimiento

Se les informó a los participantes los objetivos de la presente investigación, de igual forma se aclaró que los datos obtenidos se mantendrían de forma confidencial y serían utilizados sólo a los fines de este estudio. Asimismo, se les explicó que debían responder a preguntas múltiples opción en relación al conocimiento en Leyes. Dichos reactivos contienen tres alternativas de respuesta donde sólo una era correcta. La administración se realizó en lápiz y papel, a cada alumno se le hizo entrega del cuadernillo de preguntas y protocolo de respuesta. El tiempo estimado de administración fue de 40 minutos. Antes de comenzar con la administración de la prueba se les manifestó que su participación era voluntaria y que podía abandonar el estudio cuando lo desearan.

Análisis de datos

Los análisis se realizaron desde el Modelo de Rasch utilizando el programa Winsteps versión 3.63.2 (Linacre, 2007). Se evaluó la independencia local, el ajuste de los datos al modelo, los índices de separación y fiabilidad, el funcionamiento diferencial del ítem y la objetividad específica. El supuesto de independencia local se evaluó inspeccionando la matriz de los residuos (se esperan valores inferior a 0.025) y la matriz de varianza y covarianza (se esperan valores inferior a 0.25). Además se evaluó el ajuste de los datos al modelo a partir del ajuste de los ítems y el ajuste de las personas. Dicho análisis se realizó a partir de los valores estadísticos Infit y Outfit como medida cuadrática de los residuales (MnSq) donde valores próximos a 1 indican un ajuste perfecto entre los datos y el modelo. Según los criterios establecidos por Linacre (2002) la región para considerar un ajuste aceptable se ubica entre 0.6 y 1.3.

Respecto a los índices de separación y fiabilidad de personas e ítems, se considera adecuado un índice de separación superior a 2 (Bond & Fox, 2001) con una confiabilidad asociada de .80 (Gauggel et al., 2004). Además, se obtuvo el Mapa de Wright que permite observar en un mismo continuo la distribución de las personas y de los ítems. Asimismo se presenta gráficamente la Curva Característica del Ítem (CCI) y la Función de Información del Test (FIT). Por un lado, la "Curva Característica del Ítem" es una función matemática que establece que cada rasgo latente está en relación con la conducta que una persona manifiesta frente a un ítem (Attorresi, Lozzia, Abal, Galibert, & Aguerri, 2009), de allí que cada ítem tendrá su propia Curva Característica del Ítem (CCI) a partir del cual podremos observar su dificultad (Baker, 2001); por el otro la Función de Información (FI) de cada ítem permite estimar en qué nivel de habilidad resulta más preciso dicho reactivo; asimismo a partir de la suma de las FI de los ítems podremos obtener la FIT (Attorresi et al., 2009).

Cabe mencionar que además se examinó si existía funcionamiento diferencial de los ítems (DIF) respecto al sexo de los participantes. Un ítem presenta DIF cuando la probabilidad de respuesta correcta no depende sólo del nivel de habilidad de la persona en el rasgo intencionadamente medido por el test. Para aplicar el DIF, se realizaron análisis *pairwise* en dónde el nivel de significación se fijó en $\alpha < 0.05$, y se tuvo en cuenta que el contraste del DIF debe ser superior a ≥ 0.5 logits (Linacre, 2018).

Por último se evaluó la objetividad específica ya que para que una medida sea considerada válida y generalizable, ésta no debe depender de las condiciones específicas con que ha sido obtenida. Por lo cual, en este trabajo para analizar la invariancia de los parámetros de los ítems: (a) se dividió la base en dos de forma aleatoria, (b) se estimaron los parámetros de dificultad de los ítems y (c) se realizó el análisis de 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, 2003).

Resultados

El supuesto de independencia local se evaluó inspeccionando la matriz de los residuos y la de covarianza. Los valores observados en la matriz de covarianza no reflejaron valores superiores al valor de corte (0.25). Respecto a los resultados de la matriz residual, un 4% de los residuos presentaron valores superiores a 0.025.

Los análisis respecto al ajuste de los datos al modelo indican que en el ajuste de los ítems la medida de dificultad (δi) de los ítems varió entre $-2.42 \leq \delta i \leq 1.66$ (*M*= 0.00; *DE*= 1.03). Los valores de Infit (MnSq) de los ítems variaron entre 0.89 y 1.17, (M= 1.00; DE=0.06) y los Outfit (MnSq) 0.81 y 1.31 (*M*=1.03; *DE* = 0.11), por lo que todos los ítems se ajustan al modelo. El análisis de ajuste de las personas refleja que el modelo explicó el 88,2% de los patrones de respuesta de los sujetos (Infit y Outfit ≥ 1.3). Los niveles de habilidad variaron entre $-1.49 \leq \theta \leq 1.72$ (*M*=.13; *DE*=0.60) con valores Infit que oscilaron entre 0.64 y 1.56 (*M*=1.00; *DE*=0.18) y Outfit entre 0.58 y 1.86 (*M*=1.00; *DE*=0.25).

Respecto a los índices de separación y fiabilidad para los ítems éstos fueron adecuados (separación= 5.62; fiabilidad= .97). En relación a las personas el índice de separación fue de 1.06 y el de fiabilidad de .53, ambos índices no resultan satisfactorios. En la Figura 1 se muestra la distribución 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 este mapa se observa que la mayoría de los ítems se ubican en una posición centrada respecto a los estudiantes evaluados, lo que indica que tanto las personas como los ítems se distribuyen de manera uniforme a través de la media. Sin embargo, por ejemplo entre los LF1_52 y el LF1_1, deberían agregarse ítems con niveles de dificultad que permitan medir el nivel de habilidad que los sujetos muestran en esa parte del continuo.

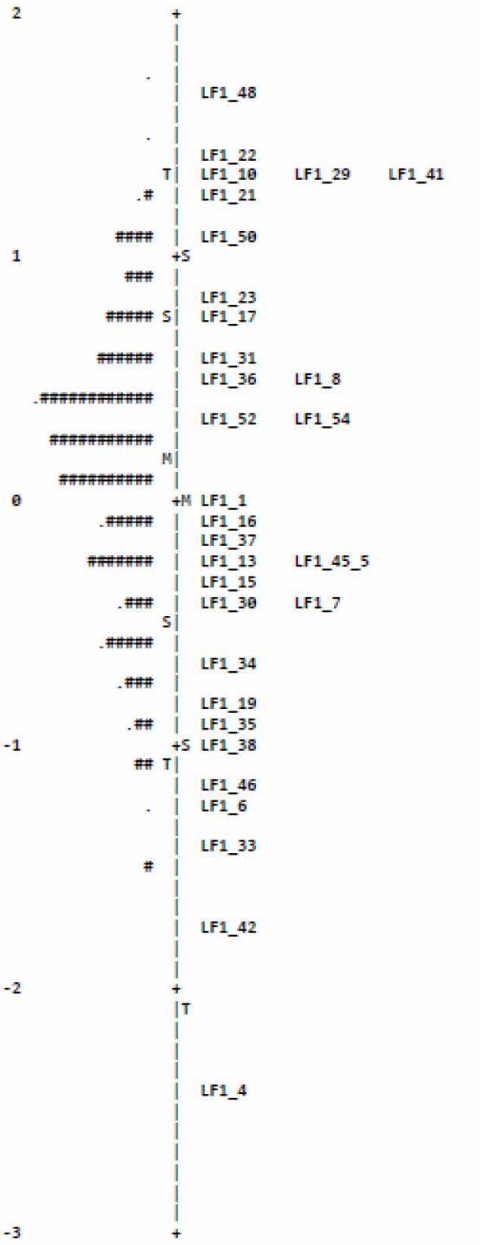

Figura 1. Mapa de personas e ítems. Del lado derecho se presenta la distribución de las personas, cada "#" significa dos personas y "." una persona. Del lado izquierdo se muestra la distribución de los ítems.

En la figura 2 se presenta la Curva Característica de cinco (5) ítems con diferentes niveles de dificultad: LF1_4, LF1_13, LF1_23, LF1_38, LF1_48. Como se observa, las líneas se trasladan desde la izquierda hacia la derecha, siendo la primera línea de la izquierda el ítem LF1_4 el de

menor nivel de dificultad, en esta curva podemos observar dónde funciona el ítem en la escala de habilidad. Si bien organizamos los ítems en orden creciente de dificultad, observamos por ejemplo que el ítem LF1_38 pudo responderlo personas con menor nivel de habilidad de lo que esperábamos respecto a este ítem, por otro lado el ítem LF1_3 presentó un comportamiento inverso.

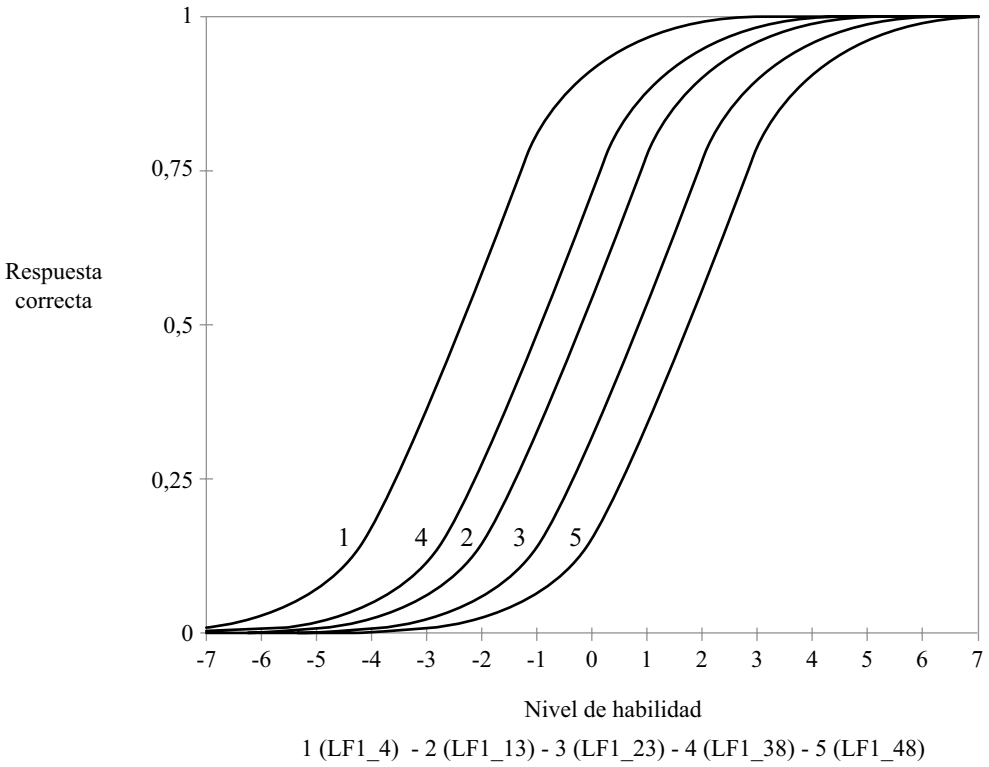

Figura 2. Curva Característica de 5 Ítem del TCG-Leyes Nivel I

Para finalizar, cabe mencionar que además de la CCI en la Figura 3 se presenta la FIT, observando el comportamiento total de los ítems. En efecto se obtiene una curva en forma de campana, siendo el pico más alto de la curva el indicador del mayor nivel de precisión del test. El pico más alto se observa en el rango de habilidad de -0.5 a 0.5 por lo que los ítems fueron suficientemente informativos de un nivel medio de habilidad.

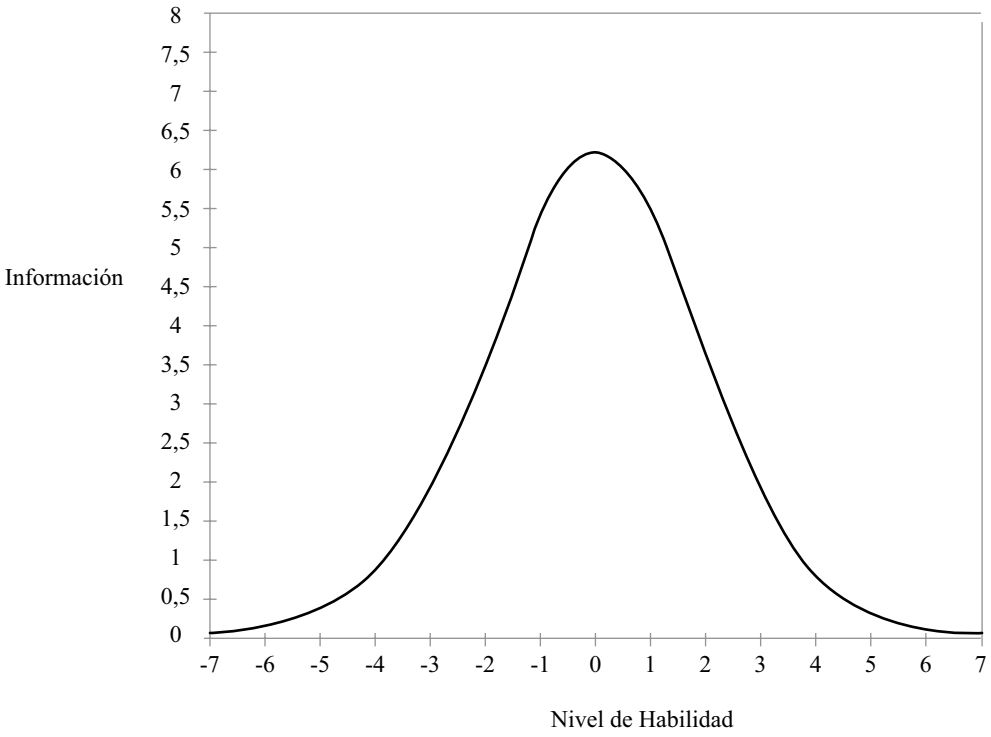

Figura 3. Función de Información del Test

Los análisis realizados respecto al *DIF* en relación al sexo, arrojaron como resultado que un solo ítem presentó un funcionamiento diferencial en relación al género. De allí que el contraste DIF para el ítem LF1_38 (Derecho constitucional) fue de 0.78 estadísticamente significativo a $p < .05$. La dificultad de los ítems (Media del DIF) para la muestra femenina fue de -1.35 logits y para la muestra masculina de -0.56 logits lo que indica que este ítem resulta más fácil para las mujeres que para los valores. En relación a la objetividad específica los resultaron mostraron un valor de r=0,946, en donde el valor de la constante fue de 0,000 y β= 0,946. De allí que se puede asumir la invariancia de los parámetros de los ítems anclas (Prieto & Delgado, 2003).

Discusión

En el ámbito educativo es imprescindible generar herramientas de medición que nos permitan evaluar el nivel de aprendizaje de los estudiantes. En la educación superior conocer cuál es el logro de los estudiantes permite obtener un indicador del nivel de formación profesional que se les está otorgando a los alumnos. Por lo cual, la construcción de evaluaciones a gran escala, con adecuadas propiedades psicométricas, resulta fundamental para establecer mediciones que nos permitan comparar resultados dentro de una institución como también entre las diversas unidades académicas. Por lo que en este trabajo, se presentan los análisis de las propiedades

psicométricas de los ítems del Test de Conocimiento en Leyes (Nivel I) mediante la aplicación del modelo de Rasch. En efecto, aquellos ítems que funcionen adecuadamente pasarán a forma parte del BI del TCG.

En líneas generales, los ítems del dominio de Leyes del Nivel I presentan propiedades psicométricas adecuadas. De los resultados obtenidos, se verifica que la respuesta que computa cada sujeto a un ítem no depende de las respuestas que dio a los otros ítems del test. Respecto al ajuste de los datos al modelo de Rasch, todos los ítems presentaron buen ajuste. En cuanto al ajuste de las personas el modelo explicó el 88,2% de los patrones de respuestas de los individuos.

Otro aspecto a mencionar refiere al índice de separación y fiabilidad, ambos índices en relación a los ítems resultó satisfactorio. Sin embargo, esto no sucedió en los índices de separación y fiabilidad de las personas. Asimismo, en el mapa de personas e ítems se observa que la mayoría de los ítems se ubican en una posición centrada, logrando una adecuada distribución de los reactivos en el continuo. Si bien la dificultad de los ítems permite medir diferentes niveles de habilidad de las personas, deberían agregarse ítems que permitan cubrir el espectro completo en nivel de dificultad para poder evaluar todos los niveles de habilidad que manifiesta la muestra. Cabe agregar, que en la conformación del instrumento se estableció que los ítems se organizarán en niveles crecientes de dificultad, sin embargo como vemos en la CCI algunos reactivos requerían mayor o menor nivel de habilidad del nivel de dificultad que se había establecido. Además, en el gráfico de la FIT vemos que el instrumento resulta preciso para medir el conocimiento en Leyes de los niveles medios de habilidad.

Con respecto al DIF, sólo para el ítem LF1_38 los resultados muestran que dicho reactivo resulta más fácil para las mujeres que para los varones. Debemos consideran que el 60% de la muestra que respondió a este instrumento es del sexo femenino y el 40% restante es de sexo masculino, por lo que en un futuro debe ampliarse la muestra y en efecto obtener una muestra de sujetos más heterogénea.

Otro punto a analizar desde los análisis psicométricos del Modelo de Rasch es la objetividad específica del test. A partir de estos resultados se logró determinar que es posible estimar los parámetros de los ítems independientemente de la muestra de sujetos que respondió al instrumento y de igual forma, determinar la habilidad de los estudiantes que contestaron sin que la medida de rasgo dependa de las características del test que respondieron.

Limitaciones

Cabe mencionar que existen ciertas limitaciones en el presente estudio. Por un lado, los análisis se realizaron a partir de una muestra pequeña de sujetos. Como bien se sabe, para aplicar TRI se necesitan muestras grandes, no obstante en la literatura existen estudios (Chen et al., 2014; Linacre, 1994) que utilizan muestras pequeñas de sujetos. De igual forma, se planifica obtener una

muestra considerablemente superior para arribar a nuevas conclusiones y poder realizar nuevos análisis psicométricos desde la TRI con el objetivo de poder calibrar un número mayor de reactivos. Además, cabe mencionar que en nuestro país nos enfrentamos a un cambio e implementación de un nuevo código Civil y Comercial. Considerando que el dominio de Leyes se basa en materias codificadas, se debería revisar el contenido de aquellos ítems relacionados con los cambios establecidos y que podrían quedar obsoletos para la formación de los nuevos estudiantes de la carrera de Abogacía. Cabe mencionar que los 31 ítems pasarán a formarán parte del Banco de Ítems del TCG, no obstante debe ampliarse la muestra de ítems y de sujetos del dominio de conocimiento en Leyes. Asimismo, deben supervisarse aquellos inconvenientes presentes, respecto a la representatividad del contenido del test y a su vez, ampliar el banco de ítems con reactivos que nos permitan cubrir el espectro completo en relación a los niveles de dificultad.

Conclusión

Partiendo de la concepción que medir los conocimientos adquiridos de los estudiantes no representa en sí mismo una mejor calidad de la educación, consideramos que la construcción de una herramienta como la presente, nos permite cualificar el nivel de logro alcanzado por los individuos en un dominio en particular y a su vez discernir quiénes han logrado adquirir las competencias y habilidades básicas en cada nivel de conocimiento. Por lo cual, tanto la enseñanza, el aprendizaje y evaluación resultan procesos que, aunque se organicen en distintos momentos, son aspectos complementarios e inseparables (Rodríguez, Muñoz, & Castillo, 2014). Por lo que este trabajo realiza aportes significativos en el ámbito de la medición y evaluación de los conocimientos y habilidades que deben adquirir los estudiantes del dominio leyesen nuestro medio.

Referencias

Arancibia, V. (1997). Los sistemas de medición y evaluación de la calidad de la educación. Santiago de Chile: UNESCO.

Attorresi, H. F., Lozzia, G. S., Abal, F. J., Galibert, M. S., & Aguerri, M. E. (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 http://www.redalyc.org/articulo.oa?id=281921792007

Baker, F. B. (2001). The basics of Item Response Theory. ERIC clearinghouse on assessment and evaluation(2nd ed). Recuperado de http://echo.edres.org:8080/irt/baker/final.pdf

Bond, T. G., & Fox, C. M. (2001). *Apllying the Rasch Model*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23* (2) 485-493. doi: 10.1007/s11136-013-0487-5

Cupani, Marcos, Ghio, Fernanda Belén, Leal, María Florencia, Giraudo, Gimena Mariel, Castro Zamparella, Tatiana, Piumatti, Gisella, Casalotti, Antonella Belén, Ramírez, Juan Claudio, Arranz, María Andrés, Farías, Analía Norma, Padilla, Natalia, & Barrionuevo, Leandro. (2016). Desarrollo de un banco de ítems para medir conocimiento en estudiantes universitarios. *Revista de psicología (Santiago), 25*(2), 1-18. doi: 10.5354/0719-0581.2017.44808

Gauggel, S., Heinemann, A. W., Böcker, M., Lämmler, G., Borchelt, M., & Steinhagen- Thiessen, E. (2004). Patient-staff agreement on Barthel index scores at admission and discharge in a sample of elderly stroke patients. *Rehabilitation Psychology*. *49*(1), 21-27. doi: 10.1037/0090-5550.49.1.21

Iraurgi, I., Lozano, O., González-Saiz, F., & Trujols, J. (2008). Valoración psicométrica de la Escala de severidad de la de-pendencia a partir de dos modelos de análisis: La Teoría Clásica de los Test y la Teoría de Respuesta al Ítem. *Boletín de Psicología*, *93*, 41-57. Recuperado de https://www.uv.es/seoane/boletin/previos/N93-3.pdf

Linacre, J. M. (1994). Sample size and item calibrations stability. *Rasch Measurement Transactions*, *7*(4), 328. Recuperado de https://www.rasch.org/rmt/rmt74m.htm

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

Linacre, J. M. (2007) Winsteps (Version 3.63.2)[*Software* de computación]. Chicago, IL: Winsteps.

Linacre, J. M. (2018). A User's Guide to Winsteps® Ministep Rasch-Model Computer Programs: Program Manual 4.2.0. Chicago: WINSTEPS.com, 2018. [Internet] Recuperado de: http://www.winsteps.com/winman/copyright.htm

Martínez Arias, R. (2006). La metodología de los estudios PISA. *Revista de Educación*, *1*, 111-129. Recuperado de http://www.revistaeducacion.mec.es/re2006/re2006_08.pdf

Martínez Rizo, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista electrónica de investigación educativa*, *11*(2), 1-18. Recuperado de https://redie.uabc.mx/redie/article/view/231

Martinez Rizo, F. (Coord.) (2015). *Las pruebas ENLACE y Excale. Un estudio de validación*. México: INNE. Recuperado de http://publicaciones.inee.edu.mx/buscadorPub/P1/C/148/P1C148.pdf

Navas, L., Sampascual, G., & Santed M. A. (2003). Predicción de las calificaciones de los estudiantes: la capacidad explicativa de la inteligencia general y de la motivación. *Revista de Psicología General y Aplicada, 56*(2), 225-237. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=760681

Olea, J., Ponsoda, V., Prieto, G. (eds.) (1999). *Tests Informatizados: fundamentos y aplicaciones*. Madrid. Pirámide.

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

Rodríguez-Garcés, C., Muñoz, S. M., & Castillo, R. V. (2014). Tests informatizados y su contribución a la acción evaluativa en educación. *Revista de Educación a Distancia*, 43, 1-17. Recuperado de http://www.um.es/ead/red/43/marlene.pdff

Shavelson R.J., Zlatkin-Troitschanskaia O., Mariño J.P. (2018) International Performance Assessment of Learning in Higher Education (iPAL): Research and Development. In: Zlatkin-Troitschanskaia O., Toepper M., Pant H., Lautenbach C., Kuhn C. (eds) Assessment of Learning Outcomes in Higher Education:Cross-National Comparisons and Perspectives (Methodology of Educational Measurement and Assessment) . Cham, CH: Springer International Publishing . doi:10.1007/978-3-319-74338-7_10

Vargas, G. M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, *31*(1), 43-63. doi:10.15517/revedu.v31i1.1252