Revista de Psicología
UNIVERSIDAD DE CHILE

2016, *25*(2), 1-18
ISSN impreso: 0716-8039
ISSN en línea: 0719-0581
www.revistapsicologia.uchile.cl

Desarrollo de un banco de ítems para medir conocimiento en estudiantes universitarios

Development of an Item Bank to Measure Knowledge in University Students

Marcos Cupani, Fernanda Belén Ghio, María Florencia Leal, Gimena Mariel Giraudo, Tatiana Castro Zamparella, Gisella Piumatti, Antonella Belén Casalotti, Juan Claudio Ramírez, María Andrés Arranz, Analía Norma Farías, Natalia Padilla, & Leandro Barrionuevo

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

Resumen: La medición en el ámbito educativo del rendimiento académico de los estudiantes universitarios es considerada empírica y cuantitativa. De allí que el propósito principal de dichas evaluaciones consiste en el control de los sistemas educativos y la evaluación a partir de criterios objetivos (Long, Wendt, & Dunne, 2011). Este trabajo apunta a desarrollar un banco de ítems para el Test de Conocimiento General compuesto de 20 dominios específicos. Se presentan avances realizados en seis dominios (psicología, biología, historia, literatura, economía y leyes). La muestra estuvo compuesta por 6.794 estudiantes. Se evaluaron 1.526 ítems de distintos dominios. Se realizó un análisis factorial exploratorio no lineal, se obtuvieron los índices de dificultad y discriminación según la teoría clásica de los test y la teoría de respuesta al ítem; también se obtuvieron índices de fiabilidad. El 68% presenta dificultad moderada y 32% un índice de dificultad alto o bajo. Sobre los índices de confiabilidad en la mayoría de los dominios se obtuvieron valores satisfactorios superiores a ,70. Se concluye la necesidad de revisar los ítems que no cumplieron estos criterios y ampliar la muestra. Este instrumento permitirá reducir los errores de clasificación de los alumnos y medir el desempeño académico con una escala de intervalo.

Palabras clave: Test de Conocimiento General, banco de ítems, teoría clásica de los test.

Abstract: Measurement in the educational field of academic achievement of university students is considered empirical and quantitative. Hence, the main purpose of such assessments is to control educational systems and evaluation based on objective criteria (Long, Wendt, & Dunne, 2011). The aim of this article was to develop an Item Bank for General Knowledge Test composed by 20 specific domains of knowledge. Considering that an effective construction of a test requires organization and systematization of activities, progress in six domains are presented. The sample was composed by 6,794 university students. 1,526 items from different domains were evaluated. To calibrate the items, a non-linear exploratory factorial analysis was performed. Difficulty and discrimination indices were obtained according to the classical theory of tests and the item response theory, and reliability indices as well. It was observed that 68% of the items have moderate difficulty and 32% of them have high or low difficulty. Internal consistency of the instrument showed high reliability values, up to .70. Further studies are needed in order to expand the item sample, and review items that showed inadequate indexes on discrimination, difficulty and reliability. This instrument allows measuring academic performance on an interval scale level and reducing the misclassification of students.

Keywords: General Knowledge Test, items bank, classical test theory.

Este trabajo ha sido financiado con subsidios de investigación y desarrollo otorgados por el Fondo para la Investigación Científica y Tecnológica de la Agencia Nacional de Promoción Científica y Tecnológica (Foncyt-PICT-2012), por el Consejo Nacional de Investigaciones Científicas y Técnicas (PIP 2012-2014), Ministerio de Ciencia y Tecnología de la Secretaría de Promoción Científica (PID 2010). Los autores agradecen la colaboración de Brenda de Dio, Daniela Denegri Coumeres, Rocío Martínez, Nilton Fernando Meza y Patricia Cataneo, por su contribución en la recolección de datos, y a los profesores de las distintas carreras universitarias que facilitaron el acceso a los estudiantes.

Contacto: M. Cupani. Cipsi - Conicet, Facultad de Psicología, Universidad Nacional de Córdoba, Ciudad Universitaria, Córdoba 5000, Argentina. Correo electrónico: marcoscup@gmail.com

Cómo citar: Cupani, M., Ghio, F. B., Leal, M. F., Giraudo, G. M., Castro Zamparella, T., Piumatti, G.,… Barrionuevo, L. (2016). Desarrollo de un banco de ítems para medir conocimiento en estudiantes universitarios. *Revista de Psicología*, *25*(2), 1-18.
http://dx.doi.org/10.5354/0719-0581.2017.44808

Introducción

Actualmente se considera que más allá de la especificidad teórica que caracteriza cada línea de pensamiento y enfoque, la evaluación de conocimiento es una dimensión constitutiva de la enseñanza y el aprendizaje (Cols, 2009). La medición en el ámbito educativo del rendimiento académico de los estudiantes universitarios es considerada empírica y cuantitativa. De allí que el propósito principal de dichas evaluaciones consiste en el control de los sistemas educativos y la evaluación a partir de criterios objetivos (Long, Wendt, & Dunne, 2011). Es así que la problemática existente en el proceso de evaluación de la educación se constituye como objeto de estudio y atención por parte de las agencias estatales, instituciones educativas, centros de investigación y comunidad educativa en general (López Jiménez & Puentes Velásquez, 2010). A lo largo del tiempo el sistema educativo ha utilizado distintas formas de evaluación para estimar el aprendizaje. De allí que entre los instrumentos más empleados para medir la adquisición de los contenidos curriculares por parte de los alumnos encontremos los exámenes de conocimiento, rendimiento y aptitudes.

Tradicionalmente, cada profesor establece en su área o materia sus propios criterios y procedimientos de calificación, particularmente en el ámbito universitario (Navas, Sampascual, & Santed, 2003). De modo que la asignación de puntuaciones está sujeta a fuentes de variabilidad no siempre atribuibles al nivel de competencia de los alumnos (Rodríguez-Ayán Mazza, 2007). A razón de ello surge el interés por desarrollar herramientas de medición correctamente elaboradas y calibradas que aporten objetividad a la evaluación de conocimiento. Contar con instrumentos estandarizados permite que los resultados sean comparables en las distintas instituciones a través de los años. Asimismo permite evaluar el cumplimiento de las metas educativas y el desarrollo de políticas educativas (Froemel, 2009).

A nivel internacional existen diferentes pruebas estandarizadas que pretenden evaluar el nivel de conocimiento adquirido por los estudiantes (Martínez Rizo, 2009). Por ejemplo, Estados Unidos utiliza medidas estandarizadas para medir el rendimiento académico a través del promedio de calificaciones (GPA, por su nombre en inglés Grade Point Average), el examen de evaluación escolástica (SAT, por su nombre en inglés Scholastic Assessment Test) y el examen del colegio americano (ACT, por su nombre en inglés American Collage Test). Dentro de ellos, el sistema GPA es una de las medidas de logros académicos más utilizada como criterio de admisión en las universidades (Volwerk & Yindal, 2012), en la validación de pruebas como el SAT y el ACT (Smits, Mellenbergh, & Vorst, 2002), como también para la selección del personal en el ámbito laboral (Kuncel, Credé, & Thomas, 2005). Sin embargo, el GPA tiene defectos como medida de rendimiento dada la variación de los planes de estudio y los currículums que hace que los promedios no sean comparables (Smits et al., 2002).

Otra prueba utilizada a nivel internacional es el Programa para la Evaluación Internacional de Alumnos (PISA, por su nombre en inglés Programme for International Student Assessment) desarrollado por la Organización para la Cooperación y el Desarrollo Económico (OCDE). Este examen permite realizar análisis comparativos al examinar el grado de preparación de los jóvenes para la vida adulta y, hasta cierto punto, la efectividad de los siste-

mas educativos (Vélaz de Medrano Ureta, 2006). La evaluación cubre las áreas de lectura, matemáticas y competencia científica, y las pruebas que se utilizan están desarrolladas desde la teoría de respuesta al ítem (TRI), específicamente desde el Modelo de Rasch.

En América Latina la diversidad y multiplicidad de experiencias en el desarrollo y propagación de sistemas nacionales de evaluación es una constante desde la década de 1980. Particularmente Chile ha sido uno de los principales referentes regionales en materia de evaluación de calidad (Lafuente, 2009). En este país existen dos pruebas que se encargan de medir las capacidades y logros educacionales en los alumnos. Una de ellas está bajo la supervisión del Sistema de medición de calidad de la educación y evalúa el logro de los Objetivos fundamentales y contenidos mínimos obligatorios (OF-CMO). Otro instrumento es la batería de Pruebas de selección universitaria (PSU), que consta de dos instrumentos obligatorios (matemática, y lenguaje y comunicación) y dos electivos (historia y ciencias sociales, y ciencias). Este último modelo de medición de las PSU tiene como objetivo seleccionar a los postulantes a las universidades del Consejo de rectores y por ello combina dos aspectos: habilidades cognitivas y contenidos curriculares (Bravo Urrutia et al., 2010).

En Argentina, a partir de la Dirección Nacional de Información y Evaluación de la Calidad Educativa, se comienza con el Sistema Nacional de Evaluación de la calidad de la educación en áreas como matemática, ciencias sociales y ciencias naturales. Este sistema tiene como objetivo brindar información sobre qué aprenden los estudiantes e identificar los factores asociados al aprendizaje. Sin embargo, dicho sistema presenta ciertas desventajas al no ofrecer un marco interpretativo de los resultados acorde con la segmentación de niveles educativos y las frecuencias de devolución; y considerando el tipo de información que se reporta, dificultan la construcción de una cultura de la evaluación en el país (Delich, Iaies, Savransky, & Galliano 2009). A su vez, debido a la variación de los contenidos a evaluar en relación con los cambios de gobierno (Larripa, 2009), los resultados no pueden compararse a través del tiempo generando la imposibilidad de mejorar los aprendizajes y la calidad educativa (Gvirtz, Larripa, & Oelsner, 2006).

Cabe añadir que la evaluación en el sistema educativo resulta útil para el control de calidad y eficacia de las políticas adoptadas al establecer el nivel de adecuación entre el plan de estudios y el aprendizaje de los estudiantes (Fuentes Navarro, 2006), como también valorar la calidad de la instrucción de los educadores. Es decir, los exámenes de rendimiento pueden proporcionar a las instituciones, la oportunidad de medir su propio progreso año a año en el cumplimiento de las normas establecidas por los organismos gubernamentales (Simner, 2000).

Este panorama nos permite reflexionar acerca de lo que se espera lograr en el ámbito educativo, lo que se hace para lograrlo, y lo que podemos hacer para mejorarlo (Fuentes Navarro, 2006). Particularmente en la Argentina la mayor parte de los sistemas de evaluación se concentran en la educación primaria y de nivel medio. Es por eso que surge el interés de contar con un banco de ítems que permita indagar sobre el conocimiento general de los estudiantes en el ámbito universitario. Evaluar las competencias con las que los estudiantes comienzan su carrera de formación, como así su trayectoria académica hasta la finalización. Esto permitiría

generar diferentes estrategias de enseñanza para maximizar la transferencia de los conocimientos. Se trata de una apuesta importante por la evaluación, lo que supone, tal como establece la Organización de Estados Iberoamericanos (1996), no solo un mero control, sino también una respuesta a la necesidad política y técnica de orientar los procesos de toma de decisiones y, de este modo, la mejora de la calidad de la educación.

A razón de lo expuesto, en la Universidad Nacional de Córdoba (UNC) se está desarrollando un banco de ítems (BI) para el Test de Conocimiento General (TCG) compuesto de 20 dominios específicos: psicología, biología, historia occidental, historia argentina, literatura, economía, leyes, estadística, química, física, matemática, arte, música, política argentina, geografía, astronomía, herramientas, tecnología, negocios y electrónica. Se estima que el BI-TCG estará conformado aproximadamente por 10.000 preguntas que midan estos dominios específicos de conocimiento (alrededor de 500 ítems por cada dominio) distribuidos entre cuatro o cinco niveles de instrucción (años cursados por los estudiantes).

Van der Linden y Glas (2000) señalan dos ventajas de los bancos de ítems. Por un lado, estos introducen flexibilidad en el campo psicológico y educativo en tanto posibilitan la construcción de test basados en las necesidades de evaluación que exigen el desarrollo de un test concreto. Por el otro, permite seleccionar ítems en relación con las características de los sujetos (Attorresi, Lozzia, Abal, Galibert, & Aguerri, 2009).

En el presente artículo se presentan los avances realizados en seis de los 20 dominios que componen el BI del TCG. Específicamente se presentan los resultados de los dominios de psicología (teorías de la personalidad, técnicas clínicas, teorías psicológicas e historia de la psicología); literatura (escritores, dramaturgos y poetas occidentales desde la época de la antigua Grecia hasta el presente, y de Argentina); leyes (contenidos sobre los principios básicos de las leyes penales, cívicas y empresariales de Argentina); historia (principales acontecimientos políticos, filosóficos y económicos en Europa de la antigua Grecia hasta la Segunda Guerra Mundial, y sobre historia argentina desde el la conquista española hasta nuestro días); biología (aspectos de la biología desde las células y organismos hasta niveles ecológicos); y economía (conceptos básicos de micro y macroeconomía).

Para la construcción efectiva del test se precisó de un enfoque organizado de las actividades a desarrollar, las que deben ser bien ejecutadas para que el test mida de manera precisa el dominio correspondiente (Downing & Haladyna, 2006), y así proporcione evidencias de validez que apoyen las inferencias que se realicen desde la puntuaciones obtenidas. Los análisis se realizaron a partir de teoría clásica de los test (TCT) y se presentan avances respecto a la TRI.

La TCT es el enfoque clásico y predominante en la construcción y análisis de los test, se ha usado desde principios del siglo XX como modelo lineal de medición adaptable a diversas situaciones y con gran éxito en pruebas de tipo cognitivas (Muñiz Fernández, 2010). Sin embargo, presenta algunas limitaciones que disminuyen la validez de los exámenes. Las principales limitaciones de la TCT son que las características del examen y las del alumno son dependientes; es decir, la habilidad del alumno se mide mediante el número de ítems respondidos correcta-

mente en el examen. Queda en evidencia el problema de la invariancia de las mediciones y de las propiedades de los instrumentos de medida.

Para subsanar esas limitaciones, la TRI intenta establecer la probabilidad de cada ítem de ser respondido correctamente. Los parámetros estimados por el modelo permiten evaluar la calidad técnica de cada uno de los ítems por separado y del instrumento como un todo, y a la vez estimar el nivel que cada examinado presenta en el constructo de interés. En conclusión este tipo de instrumento permitiría reducir al mínimo los errores de clasificación del desempeño académico de los estudiantes universitarios, ya que los indicadores o ítems empleados posibilitarían una medición objetiva del rendimiento académico, lo que aumentaría la eficiencia de la evaluación de este al obtenerse resultados precisos y transparentes.

Método

Para la construcción efectiva de un test se requiere de un enfoque sistemático y organizado de las actividades a desarrollar. Para la construcción del banco de ítem se tuvo en cuenta los doce pasos propuestos por Downing y Haladyna (2006). Estas actividades deben ser debidamente planificadas y correctamente ejecutadas a fin de producir un test que mida de manera precisa y consistente el dominio pretendido y, a su vez, que proporcione evidencias de validez que apoyen las inferencias que se realicen a partir de las puntuaciones obtenidas por el test. Además estos autores sugieren que, en la práctica, estas actividades pueden modificarse o revisarse, por lo tanto, pueden realizarse cambios en el proceso de construcción. Las actividades que se realizaron hasta el momento se mencionan a continuación.

Paso a. Análisis de contenido y tabla de especificación

Todas las pruebas de rendimiento dependen, en gran medida, de las evidencias de validez de contenido que fundamentan y apoyan las interpretaciones que se realizarán desde las puntuaciones del test. De esta manera, el objetivo esencial en esta fase es la definición operativa, semántica y sintáctica de la variable a medir, así como las facetas o dimensiones que la componen para que pueda ser medido empíricamente. A su vez, en este paso es necesario generar una tabla de especificación que ayudará a delimitar y definir el dominio de conocimiento para cada test. De este modo, dicha tabla posibilitará una planificación sistemática que permita otorgar cierto orden y orientación para la construcción de cada instrumento, para seleccionar aquellos contenidos (u objetivos) que constituyan una muestra representativa de los aprendizajes más relevantes de cada dominio en particular.

Para definir el contenido de cada dominio se analizaron los programas de formación pertenecientes a las carreras relacionadas con los dominios de conocimiento. Este material fue organizado en una planilla (Excel) por programa, año de cursado al que pertenece el programa, unidades en que se divide cada programa (contenidos generales) y temas (contenidos específicos). Luego esta información fue procesada mediante un análisis de frecuencia con el fin de visualizar de manera descendente cuáles son los contenidos más relevantes por dominio y nivel de conocimiento (aquellos con mayor frecuencia de aparición).

Cabe agregar que un grupo de expertos evaluó la representatividad de la información recabada. Por cada dominio de conocimiento participaron entre tres y seis

docentes universitarios (jueces expertos), quienes puntuaron como frecuente-infrecuente, principal-secundario o faltante los contenidos que abarcaban el listado de frecuencias extraídas del análisis realizado anteriormente. Con base en estas observaciones se incluyeron o descartaron algunos contenidos, y la información obtenida se organizó en diversos niveles de conocimiento considerando la población meta, principalmente el año de cursado de los estudiantes de cada dominio (del nivel 0 al nivel 5, variando según el dominio). Una vez seleccionados los contenidos más representativos se conformó una tabla de especificación donde se estableció redactar aproximadamente 100 preguntas por cada nivel, considerando el nivel de representatividad de los conceptos y categorías cognitivas (conocimiento, comprensión y aplicación).

Paso b. Redacción y desarrollo de los ítems

Los ítems que conforman una prueba pueden adoptar diferentes formatos, entre los que se destacan: i) preguntas abiertas (en las que se debe elaborar la respuesta); ii) completar frases (en las que se pide a la persona que complete algunos elementos de una oración); iii) de elección alternativa (si se presentan dos alternativas de respuesta entre las que la persona tiene que elegir la correcta); y iv) de elección múltiple (cuando la persona debe elegir la opción que considera correcta entre varias alternativas de respuesta). La elección múltiple, formato elegido para la construcción del TCG, es más difícil de elaborar, pero permite una evaluación más confiable, siendo un recurso importante a la hora de evaluar grupos amplios de personas (Moreno, Martínez, & Muñiz, 2004).

Para la redacción de los ítems del TCG se contó con la colaboración de profesionales de las distintas facultades de la UNC, quienes recibieron una capacitación especial sobre los procesos de construcción de test y, fundamentalmente, sobre directrices para la construcción de ítems de elección múltiple (Haladyna, Downing, & Rodríguez, 2002). De allí que respecto al enunciado se consideró que tuviera un esquema de indagación completa y que se evitara redactar la proposición base como enunciado negativo o que pudieran confundir en la elección de la respuesta correcta. Respecto a las alternativas de respuesta, se tuvo en cuenta que cada ítem tuviera una sola opción correcta; que las alternativas fueran gramaticalmente semejantes, e igualmente aceptables desde el sentido común; que se construyeran tres alternativas; que las alternativas incorrectas tuvieran el mismo grado de especificidad que la opción correcta de respuesta; y finalmente que la opción de respuesta correcta estuviera dispuesta aleatoriamente.

A cada uno de los profesionales se le entregó la tabla de especificación donde se aclaraba cuántas preguntas debía redactar por contenido (concepto). Estas preguntas fueron organizadas en fichas y a cada una se le asignó un código único de identificación, un concepto relacionado, el tipo de categoría cognitiva que evalúa, la opción correcta y una justificación de por qué cada alternativa es una opción correcta o incorrecta. También se confeccionó un espacio para categorizar el nivel de dificultad de cada uno de los ítems.

Posteriormente estas fichas fueron entregadas a jueces que evaluaron la calidad y pertinencia de los ítems. Para cumplir con tal requisito se les hizo entrega de la tabla de especificación conjuntamente con la ficha de redacción de ítem. Esto con el objetivo de que valoraran la adecuación del contenido a la población específica

según cada nivel de conocimiento para cada dominio. Los jueces calificaron las preguntas según su nivel de dificultad como fácil, mediana y difícil. Luego de la revisión por parte de los expertos el equipo de trabajo determinó los ítems que debían modificarse si los comentarios afectaban aspectos de la redacción de los ítems; o eliminarse en caso de una representación inadecuada de los contenidos a evaluar.

Paso c. Diseño, montaje y producción del test

Los ítems fueron organizados en diferentes formas con el fin de poder evaluar sus propiedades psicométricas. Para los distintos niveles de cada dominio de conocimiento se confeccionaron una forma A y una forma B, y en algunos casos, una forma C. La distribución de los ítems en cada forma se realizó por nivel de dificultad ascendente considerando los diferentes contenidos.

Además, en cada forma se establecieron ciertos ítems anclas y libres. Para la selección de los ítems anclas se consideró que respondieran a los diferentes niveles de dificultad (baja, media y alta) y que abarcaran los diferentes contenidos del nivel y dominio en particular. Por otro lado, se configuraron las formas y la cantidad de ítems a incluir estimando que los usuarios deberían poder responder el instrumento en condiciones normales, y en 40 y 60 minutos. Este criterio condicionó la cantidad de ítems anclas y libres a incluir en cada forma y dominio en particular. La respuesta correcta varió de ubicación de forma aleatoria. Asimismo se estableció un formato estándar para la conformación de cada test: a) un cuadernillo de preguntas de doble carilla para facilitar la lectura y b) un protocolo de respuesta para organizar las puntuaciones de los evaluados con espacios determinados para la elección de su respuesta (A, B o C); en este último se incluyeron ciertos datos sociodemográficos tales como edad, sexo, universidad, facultad y carrera, entre otros.

Paso d. Administración del test

La administración de la prueba se realizó a estudiantes universitarios de diferentes años de cursado de diversas carreras de la ciudad de Córdoba. Las tomas se realizaron de forma colectiva, en un horario regular de clase y bajo supervisión de los profesores asignados al horario de cursado. Previo a la administración se explicó a los estudiantes que debían responder un número de preguntas de opción múltiple, las cuales solo tenían una única opción correcta. De igual manera se sugirió que tratasen de responder a todas las preguntas y que, en caso de considerar que la pregunta era totalmente ajena a sus conocimientos, no emitiesen respuesta alguna. Luego de esta aclaración se entregó a los alumnos el consentimiento informado y el material para leer y responder.

Paso e. Análisis de datos

Para evaluar la validez de estructura interna de cada dominio se realizó un análisis factorial no lineal (AFNL). Se utilizó 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/o multidimensional del modelo ojiva normal (Ayala, 2009). NOHARM produce una matriz residual para evaluar el ajuste del modelo, dicha matriz 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), en que valores cercanos a 0 representan un ajuste adecuado al modelo. 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.

En segundo lugar, desde la TCT se realizó el análisis de ítems para determinar si el rango de dificultad y de discriminación de los reactivos era adecuado. Para estos análisis se utilizó el programa ViSta (Young, 1996). Uno de los índices más importantes para determinar el grado de dificultad de los ítems es el valor de *P*, que indica el porcentaje de la muestra que respondió de manera correcta el ítem. Por lo tanto, mientras mayor es el valor *P*, el reactivo es más fácil; un ítem con un valor *P* de ,75 indica que el 75% de todos los estudiantes de la muestra contestó el reactivo correctamente. Se consideró como criterio que los niveles de dificultad deseable para los ítems entre valores de *P* = ,30 y ,70, es decir, ni excesivamente difíciles ni fáciles (Kaplan & Saccuzzo, 2006). Se realizó una correlación de cada ítem con el puntaje total de la prueba. Este índice permite identificar la capacidad del ítem para discriminar (diferenciar) entre los individuos que poseen "más" un rasgo y los que poseen "menos" de ese rasgo. El estadístico usual es el coeficiente punto-biserial cuando las variables son dicotómica (Velandrino, 1998). Los ítems con correlaciones no significativas o bajas con el puntaje total (inferiores a ,30) deben revisarse. Para evaluar la consistencia interna de la prueba se utilizó el coeficiente Kuder Richardson 20 (KR-20), que es el más apropiado cuando se trabaja con ítems dicotómicos.

Por último, de manera complementaria se obtuvieron los índices de dificultad (b) y discriminación (a), basados en la TRI (modelo de dos parámetros), mediante el programa NOHARM. El parámetro de dificultad es el puntaje en la escala del rasgo (θ) cuya probabilidad de respuesta correcta es igual a 0,5, y se simboliza con *b*. En la práctica suele expresarse en una escala con media 0, desviación estándar 1 y rango de valores entre -3 y 3. Los valores negativos están asociados con reactivos fáciles, mientras que los valores positivos están asociados con reactivos difíciles. La capacidad discriminativa del ítem nos indica hasta qué punto un ítem puede diferenciar entre los examinados que poseen habilidades bajas y altas, en un nivel de dificultad (parámetro b) determinado del ítem. La capacidad discriminativa de un ítem se simboliza con *a* y se refleja en la inclinación o pendiente de la curva del ítem. Normalmente estos valores varían entre 0,3 y 2,5, y se consideran ítems muy discriminantes aquellos que poseen valores superiores a 1,34, moderadamente discriminante entre 0,65 y 1,33, y escasamente discriminantes los valores de 0,64 o inferiores. En el presente trabajo solo se presentan los resultados psicométricos del nivel 1 de los seis dominios por una cuestión de espacio y claridad.

Resultados

Dominio de psicología

Paso a. Para definir los contenidos del Test de Conocimiento en Psicología se seleccionaron 53 programas de estudio de diferentes carreras de la UNC: Facultad de Psicología, Arte, Derecho y Ciencias Sociales (Trabajo Social), Comunicación

Social, Filosofía y Humanidades, Ciencias Médicas, Odontología y Ciencias Económicas. Los programas recolectados se organizaron en una tabla de contenido, en la que se especificó nivel de conocimiento (del nivel 1 al nivel 5), programa de la materia, contenidos generales y específicos. Para el nivel 1 se consultaron ocho programas.

Paso b. Se redactaron 876 preguntas de los diferentes niveles. Diez jueces expertos evaluaron la pertenencia de los ítems. Los jueces calificaron las preguntas según su nivel de dificultad como fácil (19%), mediano (25%) y alto (25%). Del mismo modo los expertos consideraron que del pool inicial de ítems, un 75% son aceptados como están, un 25% se debe modificar, y 5% debería eliminarse. Por lo tanto, este dominio quedó conformado por 796 preguntas.

Paso c. Los 121 ítems del nivel 1 fueron distribuidos en tres formas (A, B y C) y organizados según nivel de dificultad. Cada forma constó de 67 ítems de los cuales 40 eran comunes a todas las formas y 27 ítems diferentes.

Paso d. Las tres formas fueron administradas a una muestra de 900 personas, 613 estudiantes de sexo femenino (68,1 %), 284 de sexo masculino (31,6 %), y tres participantes no informaron el sexo. La edad comprendida de los participantes fue entre los 18 y 67 años (M = 21,3; DT = 5,69).

Paso e. Para la forma A, el valor del RMSR (0,012) es menor al error típico de los residuos estimado (0,32) lo que nos indica que los ítems del test están midiendo una sola dimensión. Sin embargo, el índice de Tanaka de bondad de ajuste (GFI) fue de ,84, valor inferior al punto de corte recomendado (,90). Este resultado indicaría que puede haber uno o más factores que explican la varianza restante (Yen, 1993). No obstante, como se trata de una prueba que mide un factor general compuesto por factores más específicos es esperable obtener una estructura factorial compleja (Tate, 2003).

Para la forma B (RMSR = 0,012; GFI = ,85), y la forma C (RMSR = 0,014; GFI = ,86), los índices de ajuste indican que se confirma que la estructura unifactorial es viable. Con respecto a los análisis de dificultad (P) y discriminación (D), los resultados muestran (ver tabla 1) que para las forma A, B y C, se puede considerar que el 72% de los ítems presenta un nivel de dificultad moderado mientras que el 28 % restante presenta niveles muy bajos o muy altos. En relación con los índices de discriminación se observó que los valores del coeficiente punto-biserial variaron entre ,01 a ,32 para la forma A, entre ,02 a ,40 para forma B, y entre ,04 a ,47 para la forma C. También se puede observar que los índices de fiabilidad para las tres formas fueron satisfactorios, alcanzando resultados de ,77, ,77 y ,85, respectivamente para las formas A, B y C.

En los análisis de dificultad (b) y discriminación (a) desde la TRI, se observó que en la forma A los parámetros de dificultad variaron entre b = -3,50 a 2,61 y los parámetros de discriminación entre a = -0,11 a 0,79; podemos destacar que tres ítems presentaron valores negativos y deberían ser revisados o eliminados del modelo. En la forma B, los parámetros variaron entre b = -4,39 a 7,73 y a = 0,15 a 0,76 para dificultad y discriminación respectivamente; y para la forma C, entre b = -5,73 y 10,4 y a = 0,01 y 1,01. Estos resultados nos indican que los ítems presentan una variación adecuada entre los índices de dificultad, pero no así con su propiedad de discriminación.

Dominio de biología

Paso a. Se recolectaron 55 programas pertenecientes a siete carreras relacionadas con las CN y ciencias de la salud (CS) de la UNC. Cada material fue organizado por programa, año de cursado al que pertenece el programa (n = 4), unidades en que se dividen cada programa (contenidos generales) y temas (contenidos específicos). Esta información fue organizada en cuatros niveles (1 al 4). Para el nivel 1 se utilizaron 18 programas.

Paso b. Diez profesionales redactaron 532 preguntas iniciales y cinco jueces evaluaron el contenido específico seleccionado para cada ítem. Se descartaron algunas preguntas y el pool final quedó conformado por 487 ítems.

Paso c. Los 100 ítems del nivel 1 se constituyeron en dos formas (A y B), cada una de ellas con 40 ítems libres y 20 anclas.

Paso d. Las dos formas fueron administradas a una muestra de 615 personas, 387 estudiantes de sexo femenino (63 %), y 228 de sexo masculino (37 %). Las edades de los alumnos variaron entre 18 y 37 años [*M* = 21; *DT*= 3,9]).

Paso e. Los resultados del AFNL de la forma A (RMSR = 0,014; GFI = ,85) y forma B (RMSR = 0,015 y GFI = ,84) indican que la estructura unifactorial se ajusta a los datos. Con respecto a los índices de *P* y *D*, los resultados muestran que el 90% de los ítems presenta un nivel de dificultad moderado acorde a los conocimientos de la muestra; el 10% restante presenta muy baja dificultad o muy alta. Con relación a los índices de discriminación se observó que los valores del coeficiente punto-biserial variaron entre ,03 a ,49 (forma A) y entre ,00 a ,54 (forma B). Los índices de fiabilidad fueron satisfactorios con valores de ,90 y ,89, para las formas A y B, respectivamente. Los parámetros de dificultad para la forma A variaron entre *b* = -8,74 a 7,13; y los índices de discriminación variaron entre *a* = -1,26 a 0,96; y en el caso de la forma B, los parámetros de dificultad variaron entre *b* = -4,54 a 5,35 y los de discriminación entre *a* = -0,93 y 1,16.

Dominio de historia

Paso a. Se recolectaron 122 programas de formación del nivel secundario y del profesorado de Historia perteneciente a la Facultad de Filosofía y Humanidades de la UNC. Cada material fue organizado por materia, año de cursado, contenidos generales, y temas que contemplan cada unidad. Esta información fue organizada en cinco niveles (del nivel 0 al 4). Para el nivel 1 se consultaron 18 programas.

Paso b. La redacción de los ítems estuvo a cargo de dos profesores expertos en historia, a quienes se les entregó una tabla de especificación, que contenía el número de preguntas a redactar por contenido. Se redactaron 493 preguntas; luego de una revisión por tres pares expertos, el pool quedó conformado por 450 ítems. El 87% de los ítems fue aceptado, un 5% con modificaciones menores y un 8% de las preguntas fueron eliminadas.

Paso c. Para el nivel I se establecieron dos formas (A y B) de 50 preguntas cada una, 32 ítems anclas y 18 ítems libres, utilizando 68 ítems. La cantidad de ítems a incluir se realizó considerando la extensión de los reactivos y el tiempo necesario para responder la totalidad de la prueba.

Paso d. El test se aplicó a una muestra de 384 estudiantes, 234 de sexo femenino (60,9%) y 138 de sexo masculino (35,9%), un 3,1 % de la muestra no computó el se-

xo; sus edades estaban comprendidas entre 18 y 64 años (*M* = 25,01; *DT* = 14,57).

Paso e. Los resultados del AFNL de la forma A (RMSR = 0,013; GFI = ,87) y forma B (RMSR = 0,015 y GFI = ,87) indican que la estructura unifactorial se ajusta a los datos. La mayoría de los ítems presentó una dificultad moderada (63%), el 37% restante niveles de dificultad bajos o altos. El coeficiente punto-biserial fue de ,02 a ,45 para la forma A; y de ,01 a ,49 para la forma B. Los parámetros de dificultad en la forma A variaron entre *b* = -7,14 a 3,04 y los parámetros de discriminación entre *a* = -0,02 a 1,26. En la forma B, los parámetros variaron entre *b* = -4,37 a 4,70 y *a* = -0,08 a 1,06.

Dominio de literatura

Paso a. Se seleccionaron 58 programas de diferentes unidades académicas. Cada material se organizó por programa, año de cursado, unidades en que se divide cada programa y temas. Esta información se organizó en tres niveles (1 al 3). Para el nivel 1 se consultaron cuatro programas.

Paso b. Expertos en el dominio redactaron 485 preguntas sobre literatura general y argentina.

Paso c. Los 99 ítems del nivel fueron distribuidos en dos formas (A y B). Cada forma constó de 66 preguntas de las cuales 33 son anclas y 33 libres.

Paso d. Las dos formas fueron administradas a una muestra de 608 estudiantes, 426 estudiantes de sexo femenino (70 %), y 182 de sexo masculino (30 %) con edades comprendidas entre los 19 y 60 años (*M* = 24; *DE* = 6,81).

Paso e. Los resultados del AFNL indican que tanto la forma A (RMSR = 0,013; GFI = ,89) como B (RMSR = 0,014; GFI = 0,85) presentan un ajuste adecuado a los datos. En lo que respecta al índice de dificultad y discriminación (TCT) el 90% de los ítems presentan un nivel de dificultad moderado, mientras que el 10% restante presentan un nivel de dificultad muy bajo o muy alto. Los valores del coeficiente punto-biserial fueron de ,07 a ,56 para la forma A y -,00 a ,50 para la forma B. El índice de confiabilidad fue de ,89 para la forma A y ,85 para la forma B. En lo que respecta a los análisis de dificultad y discriminación desde la TRI, se observó que en la forma A los parámetros de dificultad variaron entre *b* = -3,50 a 4,09 y los parámetros de discriminación entre *a* = -0,16 a 1,06. En la forma B, los parámetros variaron entre *b* = -5,88 a 7,94 y *a* = -0,17 a 0,6 para dificultad y discriminación respectivamente.

Dominio de economía

Paso a. Se consultaron 42 programas pertenecientes a la carrera de Ciencias Económicas. Esta información fue organizada en cinco niveles (1 al 5). Siete programas fueron consultados para el nivel 1.

Paso b. Cinco profesionales redactaron 314 preguntas, que fueron sometidas a un estudio de jueces; estos recomendaron la modificación de ciertos ítems. De allí que del pool de ítems, los expertos determinaron que el 32% del total respondía a un nivel de dificultad bajo, el 55% mediana y 13% difícil. De aquellos 79% fueron aceptados sin cambios, 18% debía modificarse y el 3% eliminarse. El pool final de ítems fue de 248 ítems de los diferentes niveles de conocimiento del dominio de economía.

Paso c. Para el establecimiento de las formas del nivel 1 se seleccionaron aquellos ítems que, según las observaciones de

los expertos, eran acordes al dominio y congruentes con la población meta, como también el número de ítems a incluir se estableció a partir del tiempo necesario para responder a la prueba. Se utilizaron 100 ítems para constituir las formas A y B, cada una constó de 60 ítems (40 libres y 20 anclas).

Paso d. La muestra se formó por 603 estudiantes de la Facultad de Ciencias Económicas de la UNC, 305 mujeres (50,6%) y 260 varones (43,1%), el 6,1% no completó este dato, con edades comprendidas entre 18 y 31 años (*M* = 20,52; *DT* = 3,95).

Paso e. Los resultados del AFNL indican que tanto la forma A (RMSR= 0,013; GFI = ,92) como B (RMSR = 0,016; GFI = ,87) presentan un ajuste adecuado a los datos. El coeficiente punto-biserial presentó valores entre ,11 a ,53 para la forma A, y entre ,15 a ,55 para la forma B. En lo que respecta a la fiabilidad se obtuvieron valores de KR-20 de ,90 para la forma A y ,89 para la forma B. Por último los parámetros *a* y *b* muestran que en la forma A, los índices de *b* variaron entre -2,10 a 0,56; mientras que los índices de discriminación variaron entre 0,08 y 1,35. En la forma B los índices de *b* variaron entre -3,79 y 0,69, en lo que respecta a los valores *a*, ellos variaron entre -0,42 y 2,00.

Dominio de leyes

Paso a. Se recolectaron 36 programas pertenecientes a la Facultad de Derecho y Ciencias Sociales, de la carrera de Abogacía de la UNC. Cada material fue organizado por programa, año de cursado (n = 6), unidades (n = 305) en que se divide cada programa (contenidos generales) y temas (contenidos específicos). Esta información fue organizada en seis niveles (nivel 1 al 6). Para el nivel 1 se consultaron siete programas.

Paso b. Un profesional redactó 637 preguntas. Estas fueron sometidas a un estudio de jueces, en el cual tres expertos determinaron la pertinencia de los ítems. Del banco de ítems inicial 72% de los ítems fueron aceptados, 20% deben ser modificados y 8% eliminarse. Por lo que el pool final de ítems quedó conformado por 458 ítems de los diferentes niveles de conocimiento.

Paso c. Se utilizaron 80 ítems para constituir dos formas (A y B) con 55 ítems cada una, de los cuales 31 son ítems anclas y 24 ítems libres.

Paso d. Los test se administraron a una muestra de 170 personas, 102 de sexo femenino (60%) y 68 de sexo masculino (40%), con edades comprendidas entre los 19 y 60 años (*M* = 24,59; *DT* = 6,22), considerando un N = 85 por forma.

Paso e. El AFNL solo se realizó con los 31 ítems anclas. Los resultados indican (RMSR = 0,013; GFI = ,92) que los ítems miden una sola dimensión. El 59% de los ítems presenta un nivel de dificultad moderado y el 41% presenta niveles muy bajos o muy altos. El coeficiente punto-biserial presentó valores entre ,02 a ,42 para la forma A; y ,03 a ,51 para la forma B. Por su parte, el KR-20 arrojó un índice de ,76 para la forma A y ,54 para la forma B. Por último, el parámetro *b* presentó valores entre 5,61 y -4,45, mientras que los índices de discriminación variaron entre -0,09 a 0,73.

Tabla 1
Índices de dificultad y discriminación desde la TCT y desde la TRI de los ítems de los seis dominios

Nivel del dominio de conocimiento	Forma	N	p-valor	q-valor	Punto biserial	KR-20	b	a
Literatura 1	A	303	,07 a ,86	,14 a ,84	,07 a ,56	,89	-3,50 a 4,09	-0,16 a 1,06
	B	305	,10 a ,84	,16 a ,90	,00 a ,50	,85	-5,88 a 7,94	-0,17 a 0,6
Psicología 1	A	300	,06 a ,94	,06 a ,94	,01 a ,32	,77	-3,50 a 2,61	-0,11 a 0,79
	B	300	,07 a ,86	,14 a ,93	,02 a ,40	,77	-4,39 a 7,73	0,15 a 0,76
	C	300	,09 a ,86	,14 a ,91	,04 a ,47	,85	-5,73 y 10,4	0,01 y 1,01
Biología 1	A	304	,10 a ,94	,06 a ,90	,03 a ,49	,80	8,74 a 7,13	-1,26 a 0,96
	B	311	,18 a ,89	,11 a ,91	,00 a ,54	,81	-4,54 a 5,35	-0,93 a 1,16
Historia 0	A	306	,11 a ,97	,03 a ,89	,01 a ,48	,85	-9,68 a 7,38	0,06 a 0,90
	B	312	,23 a ,95	,05 a ,77	,06 a ,48	,82	-5,26 y 4,38	0,08 a 0,86
Historia 1	A	192	,14 a ,95	,05 a ,86	,02 a ,45	,74	-7,14 a 3,04	-0,02 a 1,26
	B	192	,10 a ,95	,05 a ,90	,01 a ,49	,75	-4,37 a 4,70	-0,08 a 1,06
Leyes 1	A	74	,16 a ,97	,03 a ,84	,02 a ,42	,76	5,61 y 4,45	0,09 a 0,73
Economía 1	A	299	,31 a ,91	,09 a ,69	,11 a ,53	,90	-2,10 a 0,56	0,08 a 1,35
	B	330	,27 a ,93	,07 a ,73	,15 a ,55	,89	-3,79 y 0,69	-0,42 y 2,00

Nota: TCT = teoría clásica de los tests; TRI = teoría de respuesta al ítem; p = proporción de respuestas correctas; q = proporción de respuestas incorrectas; Punto biserial = índice de discriminación; K-20 = índice de fiabilidad; b: índice de dificultad y a: índice de discriminación.

Discusión y conclusiones

El rendimiento académico del estudiante universitario constituye un factor imprescindible en el abordaje de la calidad de la educación superior, debido a que es un indicador que permite una aproximación a la realidad educativa (Garbanzo Vargas, 2007). La evaluación de la educación surgió como respuesta a una necesidad percibida por muchos países. Actualmente en nuestro medio, las evaluaciones destinadas a la valoración del sistema educativo no pueden ser aplicadas en distintas instituciones académicas, ya que las mismas presentan la limitación de no tener en cuenta la variación y diversidad de los contenidos que conforman el currículum.

Debido a esto, es de suma importancia contar con un instrumento estandarizado que responda a la especificidad de cada unidad académica y que, por lo tanto, contemple dichas diferencias. A razón de ello en la UNC, se está construyendo un Test de Conocimiento General que busca evaluar el conocimiento de los estudiantes en diferentes dominios de conocimiento. Contar con este amplio sistema de evaluación, permitiría saber el nivel de conocimiento con el que ingresan los estudiantes, cómo evolucionan y con qué nivel finalizan. Además, contar con un banco de ítems en distintos dominios de conocimiento introduce flexibilidad en la evaluación en el campo psicológico y educativo, ya que posibilita la construcción de test basándose únicamente en consideraciones prácticas de carácter específico, relacionadas con las necesidades de evaluación que, en un momento determinado, exigen el desarrollo de un test concreto. La segunda ventaja tiene

que ver con el uso eficiente de las respuestas de los sujetos a los ítems: cualquier conjunto de datos se puede incorporar al sistema para una actualización periódica de las estimaciones de los parámetros de los ítems (Van der Linden & Glas, 2000).

Para evaluar la adecuación de los ítems, desde la teoría clásica de los test se evaluó la calidad de las respuestas de los sujetos a los ítems y del total del test. Se observó que de los 1.526 ítems distribuidos en seis dominios (psicología, biología, leyes, economía, literatura e historia), 68% presenta dificultad moderada y el 32% restante un índice de dificultad alto o bajo. En lo que respecta a los índices de confiabilidad en la mayoría de los dominios se obtuvieron valores satisfactorios superiores a ,70, a excepción del nivel 1 del dominio de leyes (forma B) y del nivel 3 del dominio de biología (forma C).

De los resultados obtenidos se concluye la necesidad de revisar los ítems que no cumplieron estos criterios y de ampliar la muestra de los ítems. Se ha podido identificar algunos inconvenientes en la representatividad del contenido del test. Los ítems redactados no cubren todo el dominio de interés, por lo cual, se planifica ampliar el banco de ítems con preguntas de los niveles de dificultad extremos para de esta manera poder discriminar entre buenos y malos desempeños.

Por su parte, los resultados obtenidos desde la TRI permitieron superar algunas limitaciones de la TCT, ya que la primera se interesa más en las propiedades de los ítems individuales que en las propiedades globales del test. Puede decirse que uno de los supuestos fundamentales de la teoría se cumple, a saber, la mayoría de los ítems miden solo una aptitud o rasgo (unidimensionalidad). En todos los dominios el RMSR es menor al error típico de los residuos estimados de lo que se entiende que el modelo se ajusta. Sin embargo, el índice de Tanaka de bondad de ajuste fue, para algunos dominios, inferior al punto de corte recomendado (,90); este resultado indicaría que puede haber uno o más factores que explican la varianza restante (Yen, 1993).

No obstante, como se trata de una prueba que mide un factor general compuesto por factores más específicos es esperable obtener una estructura factorial compleja (Tate, 2003). A futuro, se planifica la revisión de aquellos ítems que no se ajustaron al modelo mediante nuevos estudios de expertos en el área; igualmente se considera que los resultados obtenidos para los ítems del nivel I son alentadores.

Se proyecta completar los análisis de los ítems de todos los dominios desde la TRI, Modelo de Rasch. Ya que, a saber, aunque en principio tanto la TCT como la TRI pueden trabajar con bancos de ítems, la TCT presenta limitaciones. Pues bien, dado que en la TCT los parámetros de los ítems dependen de la muestra de sujetos que ha sido utilizada para estimarlos, es difícil conseguir que los valores estimados para los parámetros de todos los ítems sean estrictamente comparables.

Por el contrario, la invarianza de los parámetros del ítem en la TRI convierte a esta teoría en el marco adecuado para trabajar con bancos de ítems, ya que permite disponer de una escala común para los parámetros de todos los ítems. En la aplicación de la TRI un paso insoslayable es optar por un modelo teórico que suministre una buena representación del rendimiento de los ítems. Dentro de ellos, el Modelo de Rasch, de un parámetro, plantea que la probabilidad de acertar un ítem depende solamente del nivel de dificultad de dicho ítem y del nivel del individuo en la varia-

ble medida. El modelo de Rasch presenta ventajas fundamentales que hace que sea ampliamente utilizado en la validación de pruebas educativas. En particular los beneficios de dicho modelo para el análisis de pruebas educativas pueden aplicarse a las pruebas PISA, a las pruebas de diagnóstico o bien a pruebas de certificación (Montero, Rojas, & Zamora, 2014).

También se planifica utilizar test adaptativos informatizados (TAI), lo que propiciaría minimizar el error estándar de medición y la posibilidad de medidas de longitud sin pérdida de precisión y fiabilidad, mejorando la posibilidad de diagnóstico con evaluaciones más breves y precisas (Olea & Ponsoda, 2003). Esto ayudaría a realizar un seguimiento longitudinal del conocimiento de un alumno, generar un diagnóstico de la cantidad y calidad de contenido adquirido, especificar qué contenido teórico dado resulta más dificultoso e incorporar nuevas alternativas de aprendizaje.

Entonces, entre los beneficios que ofrece la construcción de este instrumento se encuentra la adecuación del plan de estudios a los requerimientos y necesidades de los estudiantes (Fuentes Navarro, 2006). Es decir, la enseñanza se vería favorecida si los contenidos y la dificultad de la instrucción fueran acordes al conocimiento y habilidades del sujeto, optimizando el proceso de enseñanza (Rolfhus & Ackerman, 1999).

Asimismo, dicha evaluación posibilitaría la valoración de calidad de la instrucción de los educadores. Contar con herramientas de medición correctamente elaboradas representaría un avance en la evaluación del aprendizaje de los sistemas educativos. En conclusión, el aporte de este trabajo es significativo en el campo de la medición y evaluación en nuestro medio. El presente proyecto permitiría mejorar las trayectorias académicas, el desempeño académico y disminuir la deserción universitaria.

Referencias

Attorresi, H. F., Lozzia, G. S., Abal, F. J. P., 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

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, New York: The Guilford Press. Recuperado de http://psycnet.apa.org/psycinfo/2009-01904-000

Bravo Urrutia, D., Bosch Cartagena, M. A., Del Pino Manresa, G., Donoso Retamales, G., Manzi Astudillo, J., Martínez Martínez, M., & Pizarro Sánchez, R. (2010). *Validez diferencial y sesgo de predictividad de las pruebas de admisión a las universidades chilenas*. Santiago, Chile: CTA-PSU. Recuperado de https://is.gd/zv0Dkm

Cols, E. (2009). Introducción. La evaluación de los aprendizajes como objeto de estudio y campo de prácticas. *Archivos de Ciencias de la Educación*, *3*(3), 11-14. Recuperado de http://www.memoria.fahce.unlp.edu.ar/art_revistas/pr.4079/pr.4079.pdf

Cupani, M., Zalazar-Jaime, M. F., Garrido, S., Gross, M., & Tavella, J. (Octubre, 2012). *Construcción de un test de conocimiento general*. Trabajo presentado en el X Congreso Latinoamericano de Sociedades de Estadística, Córdoba, Argentina.

Delich, A., Iaies, G., Savransky, N., & Galliano, M. (2009). *Hacia un nuevo debate de los resultados de las evaluaciones de calidad educativa en la Argentina*. Buenos Aires, Argentina: Centro de estudios en Políticas Públicas. Recuperado de https://is.gd/cK8UWp

Downing, S. M. & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Froemel, J. E. (2009). La efectividad y la eficacia de las mediciones estandarizadas y de las evaluaciones en educación. *Revista Iberoamericana de Evaluación Educativa, 2*(1), 10-28. Recuperado de http://www.rinace.net/riee/numeros/vol2-num1/art1.pdf

Fuentes Navarro, R. (2006). La constitución científica del campo académico de la comunicación en México y en Brasil: análisis comparativo. *Revista Latinoamericana de Ciencias de la Comunicación, 5*, 48-55. Recuperado de http://www.eca.usp.br/associa/alaic/revista/r5/art_04.pdf

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

Gvirtz, S., Larripa, S., & Oelsner, V. (2006). Problemas técnicos y usos políticos de las evaluaciones nacionales en el sistema educativo argentino. *Archivos Analíticos de Políticas Educativas, 14*(18), 1-24. Recuperado de http://www.redalyc.org/articulo.oa?id=275020543018

Haladyna, T. M., Downing, S. M., & Rodríguez, M. C. (2002). A review of multiple-choice item writing guidelines. *Applied Measurement in Education, 15*(3), 309-334. http://dx.doi.org/10.1207/S15324818AME1503_5

Kaplan, R. M. & Saccuzzo, D. P. (2006). *Pruebas psicológicas: principios, aplicaciones y temas*. México, Distrito Federal., México: Thomson.

Kuncel, N. R., Credé, M., & Thomas, L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*(1), 63-82. Recuperado de http://people.uncw.edu/caropresoe/EDN523/article.pdf

Lafuente, M. (2009). La experiencia del sistema nacional de evaluación del proceso educativo, SNEPE, en Paraguay: aprendizajes y desafíos. *Revista Iberoamericana de Evaluación Educativa, 2*(1), 49-73. Recuperado de http://hdl.handle.net/10486/661545

Larripa, S. (2009). Reflexiones sobre las funciones de los sistemas de evaluación educativa de gran escala. *Archivos de Ciencias de la Educación, 3*(3), 69-78. Recuperado de http://www.memoria.fahce.unlp.edu.ar/art_revistas/pr.4083/pr.4083.pdf

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

López Jiménez, N. E. & Puentes Velásquez, A. V. (Septiembre, 2010). La evaluación de la calidad de la educación en Colombia. Estado del Arte. Trabajo presentado en el Congreso Iberoamericano de Educación, Buenos Aires, Argentina. Recuperado de https://is.gd/p6ISFc

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://is.gd/qDyHTq

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

Montero, E., Rojas, S., & Zamora, E. (2014). Quinto informe del estado de la educación. Costa Rica: Conare. Recuperado de https://is.gd/57Hj0M

Moreno, R., Martínez, R. J., & Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, *16*(3), 490-497.

Muñiz Fernández, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo, 31*(1), 57-66. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=3150824

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. (2003). *Test adaptativos informatizados*. Madrid, España: UNED. Recuperado de https://is.gd/M94CiT

Tiana, A. y Santángelo, H. (1994). Evaluación de la calidad de la educación. *Revista Iberoamericana de Educación*, *10*. VII Reunión Ordinaria de la Asamblea General de la OEI, Octubre 1994. Buenos Aires. Recuperado de http://www.rieoei.org/oeivirt/rie10a09.htm

Rolfhus, E. L. & Ackerman, P. L. (1999). Assessing individual differences in knowledge: knowledge, intelligence, and related traits. *Journal of Educational Psychology*, *91*(3), 511-526. http://dx.doi.org/10.1037/0022-0663.91.3.511

Rodríguez-Ayán Mazza, M. N. (2007). *Análisis multivariado del desempeño académico de estudiantes universitarios de química* (Tesis doctoral, Universidad Autónoma de Madrid, Madrid, España). Recuperada de https://repositorio.uam.es/bitstream/handle/10486/1800/5491_rodriguez_ayan.pdf

Simner, M. L. (2000). A joint position statement by the Canadian Psychological Association and the Canadian Association of School Psychologist on the Canadian press coverage of the province-wide achievement test results. *Canadian Journal of School Psychology*, *16*(1), 1-14. Recuperado de https://is.gd/DGrjuL

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002) Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement*, *39*(3), 187-206. Recuperado de http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2002.tb01173.x/abstract

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.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*(3), 159-203. http://dx.doi.org/10.1177/0146621603027003001

Van der Linden, W. J. & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*(1), 35-53. http://dx.doi.org/10.1207/s15324818ame1301_2

Velandrino, A. (1998). *Análisis de datos en ciencias sociales*. Murcia, España: DM Editora.

Vélaz de Medrano Ureta, C. (2006). Presentación. Una visión integral de las evaluaciones del PISA (OCDE) con especial atención a la participación en España [Edición extraordinaria]. *Revista de Educación*, 13-18. Recuperado de http://www.revistaeducacion.mec.es/re2006/re2006.pdf

Volwerk J. J. & Yindal, G. (2012). Documenting student performance: An alternative to the traditional calculation of grade point averages. *Journal of College Admission*, *216*, 16-23. Recuperado de http://files.eric.ed.gov/fulltext/EJ992990.pdf

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-214. http://dx.doi.org/10.1111/j.1745-3984.1993.tb00423.x

Young, F. W. & Bann, C. M. (1996). ViSta: The visual statistics system. *Research Memorandum*, *94*(1), 1-13. Recuperado de http://147.156.1.4/~prodat/ViSta/vista-frames/pdf/YoungBann.pdf

Fecha de recepción: 20 de septiembre de 2016
Fecha de aceptación: 2 de diciembre de 2016