2019, Vol. 19, No. 2
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

Proceso de construcción de pruebas educativas: El caso de la Prueba de Habilidades Cuantitativas

Educational tests development process: The case of the Quantitative Abilities Test

Luis Rojas-Torres * [1], Graciela Ordóñez-Gutiérrez [2]

1 - Universidad de Costa Rica, Costa Rica.

Introducción
Construcción de una prueba
Etapas previas
PHC 2018
Discusión
Referencias

Recibido: 08/05/2019 **Revisado:** 17/05/2019 **Aceptado:** 24/05/2019

Resumen

La finalidad de este artículo es brindar una guía teórica y práctica de cómo construir una prueba educativa. En la primera parte del artículo se presenta una exposición detallada de las etapas que se deben seguir para construir una prueba educativa escrita. En la segunda parte, se muestra cómo se aplicó cada una de estas etapas a la construcción de la Prueba de Habilidades Cuantitativas de la Universidad de Costa Rica, una prueba educativa con ítems de selección única que utiliza el modelo de medición de Teoría de Respuesta al Ítem de dos parámetros. A partir de la exposición, se concluye que la construcción de pruebas es un proceso riguroso, por lo que una elaboración deficiente provocaría una generación de inferencias erróneas de las habilidades de los sujetos.

Palabras clave: *prueba educativa, construcción de pruebas, evidencias de validez, Prueba de Habilidades Cuantitativas, Teoría de Respuesta al Ítem*

Summary

The goal of this paper is to present a theoretical and practical guide on how to develop an educational test. In the first part of this paper, a detailed explanation of each stage used in development of a written educational test is presented. In the second part, it is shown how this process was applied to the development of the Quantitative Abilities Test of the University of Costa Rica, which is an educational test with multiple choice items that uses the two parameter measurement model of the Item Response Theory. From this exposition, it is concluded that the test development is a rigorous process in which even a little mistake can cause wrong inferences about subjects' abilities.

Keywords: *educational test, test development, validity evidences, Quantitative Abilities Test, Item Response Theory*

***Correspondencia a:** Luis Rojas-Torres, E-mail: luismiguel.rojas@ucr.ac.cr

Cómo citar este artículo: Rojas-Torres, L., & Ordóñez-Gutierrez, G. (2019). Proceso de construcción de pruebas educativas: El caso de la Prueba de Habilidades Cuantitativas. *Revista Evaluar, 19*(2), 15-29. Recuperado de https://revistas.unc.edu.ar/index.php/revaluar/index

Introducción

Las pruebas o tests son instrumentos o dispositivos de evaluación de un dominio específico que permiten medir el grado de acierto de las respuestas que los sujetos otorgan a un conjunto de preguntas (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014; Castillo-Arredondo & Cabrerizo-Diago, 2010); además, son los instrumentos de evaluación más utilizados en los sistemas educativos, ya que permiten la valoración de componentes educativos de manera directa; asimismo, los tests posibilitan calificaciones estandarizadas a un corto plazo.

En particular, las pruebas educativas son aquellas que evalúan elementos relacionados con el proceso de enseñanza y aprendizaje de los estudiantes, como el dominio de un contenido o procesos de razonamiento en los que se recurre al contenido aprendido. Los principales propósitos con que se usan estas pruebas son: a) para hacer inferencias de los resultados educativos estudiantiles; b) para tomar decisiones con respecto a los estudiantes (certificaciones, diagnóstico, selección o ingreso a programas educativos); c) para realizar inferencias sobre los procesos de enseñanza y aprendizaje (AERA, APA, & NCME, 2014) y d) para evaluar la calidad de los sistemas educativos en pro de la mejora de estos (Castillo-Arredondo & Cabrerizo-Diago, 2010; Tiana, 1996). A pesar de que las pruebas educativas son ampliamente utilizadas en diversas áreas de la educación, su construcción muchas veces se realiza sin la rigurosidad técnica demandada por las múltiples instancias especializadas en evaluación y medición educativa y las consideraciones establecidas en los múltiples manuales que orientan su desarrollo. Por este motivo, este artículo tiene como objetivo brindar una guía teórica y práctica de cómo construir una prueba educativa escrita. Para este objetivo se presentan las etapas que rigen la construcción de una prueba educativa, de acuerdo con lo establecido en la literatura; luego, se presenta un ejemplo de la aplicación de esta guía en la construcción de la Prueba de Habilidades Cuantitativas (PHC) de la Universidad de Costa Rica (UCR), utilizada en el proceso de admisión del año 2018.

Es importante mencionar que el proceso que se describe en este artículo es aplicable tanto para pruebas estandarizadas como para "pruebas de aula" (aquellas que utilizan los docentes para la evaluación del avance de los estudiantes durante los ciclos lectivos). No obstante, la descripción de los procesos de este documento se concentra en las evaluaciones estandarizadas, las cuales disponen de un tiempo de construcción más amplio que el que tienen las pruebas de aula. La recomendación para el docente el aula es que procure acercarse lo más que pueda a los estándares descritos en este documento.

Construcción de una prueba educativa

Son varios los investigadores y especialistas en construcción de pruebas (Downing, 2006; Embretson, 2017; Ferrara, Lai, Reilly, & Nichols, 2017; Kane, 2013; Muñiz & Fonseca-Pedrero, 2019; Villareal-Galera, Alfaro-Rojas, & Brizuela-Rodríguez, 2015) que indican que para elaborar un test se requiere una secuencia de etapas, pasos o fases, de tal manera que se alcancen evidencias de validez y confiabilidad sobre las puntuaciones obtenidas en estos. A continuación se presenta una explicación de las etapas fundamentales para la construcción de tests.

Plan general

La primera etapa en la construcción de una prueba es definir el *qué* y *para qué* se quiere medir. El *qué* medir determina el constructo, lo cual implica efectuar un abordaje teórico de este y una explicación de cómo la medición del constructo permite el alcance del objetivo establecido o el propósito de la medición. Hay objetivos que se explican directamente, como la obtención de los niveles de conocimiento en el constructo de los miembros de una población, pero hay otros que son más complejos como es el uso de una prueba de razonamiento con figuras para la selección de los estudiantes de una carrera universitaria. El *para qué* medir implica establecer una explicación detallada y precisa de las razones por las cuales se elabora el test (Muñiz & Fonseca-Pedrero, 2019). Además, se debe especificar el contexto en el cual se aplicará la prueba, el cual incluye la población a la que se medirá y las circunstancias de la aplicación. También es importante tomar en cuenta qué decisiones se tomarán con respeto a las personas de acuerdo con las puntuaciones obtenidas, ya que la calificación en un test puede servir para propósitos varios como seleccionar, diagnosticar, clasificar, entre otros.

Las respuestas al *qué* y *para qué* medir determinarán todos los elementos que se considerarán en la construcción. De esta forma los reactivos de la prueba serán de respuesta cerrada si se pretende evaluar el producto final de un proceso, o serán de respuesta abierta si es para evaluar todo el proceso (Castillo-Arredondo & Cabrerizo-Diago, 2010; Mateo & Martínez, 2008). Además, el modelo de interpretación de los puntajes será con base en normas si se pretende comparar a los sujetos con el resto de la población, o será con base en criterios si lo que se quiere es establecer estándares de dominio (Martínez-Arias, Hernández-Lloreda, & Hernández-Lloreda, 2006). Por otro lado, el modelo de medición será la Teoría Clásica de los Tests (TCT), la Teoría de Respuesta de los Ítems (TRI) u otro, según las propiedades del modelo que beneficien al objetivo de medición (Muñiz & Fonseca-Pedrero, 2019). Por ejemplo, si la finalidad es maximizar la discriminación en un nivel de habilidad se puede recurrir a la TRI; pero, si se quiere medir un constructo en una población pequeña se puede recurrir a la TCT.

Cabe resaltar que el plan general es indispensable para alcanzar un grado aceptable de validez de los usos de las puntuaciones de un test, donde *validez* se entiende como "el grado en que la evidencia empírica y las justificaciones teóricas apoyan la pertinencia de las acciones e interpretaciones de las puntuaciones de las pruebas" (Messick, 1989, p. 6). Si una prueba se construye sin tener en cuenta cuáles interpretaciones se harán con sus puntajes o qué usos se les darán a sus puntuaciones, difícilmente tendrá interpretaciones o usos válidos desde la perspectiva de la rigurosidad de la evaluación y medición educativa.

Definición del contenido

En esta etapa se establece cuáles son los elementos, componentes o dimensiones del constructo a evaluar. La definición de estos componentes se realiza a partir de la revisión teórica sobre el constructo o mediante la construcción de una teoría acerca del mismo (Embretson, 2017; Muñiz & Fonseca-Pedrero, 2019). Esto con la finalidad de constituir una definición operativa del constructo y lograr obtener medidas de manera empírica. En esta etapa se empieza el desarrollo de una de las primeras fuentes de evidencias de validez de los usos de las puntuaciones: evidencias de validez basadas en el contenido, las cuales buscan que todos los elementos relevantes del constructo sean considerados (AERA, APA, & NCME, 2014).

Especificaciones del test

En esta etapa se define cómo deben ser las características del instrumento con el que se evaluará el constructo pretendido, por lo que se debe construir una tabla de especificaciones, ya que esta es un elemento indispensable a la hora de elaborar los ítems del test. En la tabla se asigna el puntaje que se debe otorgar a cada combinación de categorías de distintos aspectos del test como procesos, contenidos, dificultades, entre otros. El uso riguroso de la tabla de especificaciones permitirá que la prueba presente evidencias de validez basadas en el contenido.

Además, se debe decidir cuál será el formato de los ítems con los que se evaluarán los elementos de la tabla de especificaciones, esto es: el tipo de ítem, la longitud y el tipo de alternativas que se va a utilizar (Martínez-Arias et al., 2006; Muñiz & Fonseca-Pedrero, 2019). Igualmente, se debe dilucidar qué tiempo será necesario para resolver la prueba, los materiales que se utilizarán para la evaluación y los horarios en que se aplicará la prueba. Cada uno de estos elementos se selecciona considerando la finalidad de emplear una prueba. Por otro lado, esta etapa es la base para alcanzar evidencias de validez basadas en la estructura interna, que son aquellas evidencias de que las dimensiones establecidas en la definición del constructo se reproducen en los datos de la prueba.

Construcción de los ítems

De acuerdo con Muñiz y Fonseca-Pedrero (2019), la construcción de los ítems constituye una de las fases más importantes en la confección de un instrumento de medición, particularmente en la elaboración de una prueba educativa; puesto que los reactivos son los que conforman el instrumento. Así, una construcción deficiente incidirá en las propiedades métricas del test, lo que repercutirá mucho en las inferencias que se realicen sobre las puntuaciones que se obtengan. En este sentido, para la elaboración de los ítems es indispensable capacitar a las personas que realizarán esta labor. Esto demanda generar el perfil de los constructores, contactar a personas con ese perfil dispuestas a colaborar en la construcción, brindar una capacitación detallada de los elementos que se desean evaluar en los ítems, asignar tareas específicas de cómo y qué incluir en la elaboración de los reactivos. Luego se debe evaluar la construcción. Con base en esta evaluación se selecciona a los mejores constructores y se procede a la construcción de los ítems. Para esta etapa es recomendable otorgar y asignar los elementos particulares de la tabla de especificaciones a cada constructor.

Una vez construidos los reactivos, estos deben ser evaluados por un grupo de expertos en el constructo a medir. En esta valoración se debe analizar si los elementos pretendidos del constructo están considerados y, también, se debe indagar si no hay fuentes de varianza irrelevante al constructo, es decir, que en los ítems no se evalúen elementos que no forman parte del constructo en cuestión (Messick, 1989). Además, se debe analizar si hay fuentes de dificultad diferenciales por grupos relevantes de población. Por ejemplo, en una prueba realizada por dos culturas distintas no se deben agregar contextos familiares para una sola de ellas. En este mismo sentido, se solicita a los expertos que juzguen si los ítems cumplen con los principios básicos que deben regir la construcción de ítems, estos son de acuerdo con Muñiz y Fonseca-Pedrero (2019): representatividad, relevancia, diversidad, claridad y sencillez. Los reactivos que se aprueban con este juzgamiento pueden ser utilizados en el ensamblaje de la prueba. Es importante resaltar que en esta etapa

se rechazan muchos reactivos, sobre todo cuando los constructores son novatos, debido a esto es importante efectuar una construcción de al menos el doble de la cantidad pretendida.

Estudio piloto de los ítems

Luego de la construcción de los reactivos, se procede a ensamblar uno o varios formularios de aplicación según las características establecidas en la tabla de especificaciones. Estos formularios serán aplicados a una población con características semejantes a la población a la que va dirigida la prueba con el fin de obtener una aproximación de las propiedades psicométricas del instrumento en la población meta, lo cual permitirá determinar cuáles ítems son aptos para ser utilizados en el instrumento final (Castillo-Arredondo & Cabrerizo-Diago, 2010; Mateo & Martínez, 2008; Muñiz & Fonseca-Pedrero, 2019; Villarreal-Galera et al., 2015). En esta etapa se debe procurar mantener las mismas condiciones de administración de la prueba. En cuanto a los análisis estadísticos del test, se deben realizar los requeridos según el modelo de medición seleccionado inicialmente. La explicación de la aplicación y el análisis de ítems, se especifica en las etapas de la construcción de la prueba definitiva.

Con los datos de esta aplicación piloto, se pueden indagar evidencias de validez basadas en la relación con otras variables (Embretson, 2017; Martínez-Arias et al., 2006; Villarreal-Galera et al., 2015), que es la comprobación de relaciones teóricas del constructo pretendido con variables externas. Esto se refiere a los patrones de relación de las puntuaciones de la prueba con otras puntuaciones de rasgos y criterios empíricos que estén relacionadas con los rasgos representados por la calificación obtenida por los examinados en la prueba; por ejemplo, lugar de procedencia escolar, medidas de motivación, entre otras. Es importante que la evidencia empírica sobre las relaciones con otras variables sea consistente con los objetivos de medición para respaldar las evidencias de contenido (Embretson, 2017).

Para obtener evidencias de relación con otras variables se requiere construir una base de datos que incluya información relevante sobre los examinados, por ejemplo: cantidad de materias matriculadas, promedios obtenidos en cursos relevantes, etc. Por otro lado, se puede indagar sobre las evidencias de validez basadas en la estructura interna mediante la comprobación de la estructura factorial propuesta (Martínez-Arias et al., 2006; Mateo & Martínez, 2008). Las indagaciones de evidencias de validez desde el pilotaje permitirán que la prueba final presente las evidencias de validez requeridas.

Ensamblaje de la prueba

En esta etapa se seleccionan los ítems que serán utilizados en el examen. Para la selección de los ítems se deben seguir las condiciones establecidas en la tabla de especificaciones. Además, si en alguna de las condiciones establecidas hay un excedente de ítems, lo recomendable es seleccionar los ítems que, según los jueces, sean más pertinentes para la evaluación del constructo pretendido.

Posteriormente, se procede a generar el formulario de examen. Este formulario debe iniciar con unas instrucciones generales que señalen: a) el tiempo requerido para resolver la prueba completa; b) la forma en que debe resolverse; y, c) la estructura de la prueba en cuanto a cantidad de ítems. Luego del ensamblaje, se debe realizar una revisión detallada del formulario para garantizar que los ítems no tengan errores de forma ni de fondo.

Aplicación de la prueba

La planificación de la aplicación de la prueba es la etapa en la que se definen las condiciones necesarias para la administración del test en la población meta. Es por esto que esta etapa demanda efectuar una logística rigurosa para asegurar dichas condiciones. Un punto importante es la capacitación de los aplicadores, quienes son los encargados de administrar el test a los examinados.

La capacitación implica brindar la información sobre las labores importantes en la administración de la prueba; por ejemplo: el resguardo del material, la revisión del aula, la organización de entrada de los postulantes a las instalaciones, la explicación correcta de las instrucciones, la supervisión de la aplicación y la devolución del material. En la capacitación se debe asegurar que los aplicadores puedan brindar las condiciones requeridas para que los sujetos demuestren su verdadero nivel de habilidad en la prueba (Muñiz & Fonseca-Pedrero, 2019). Un error puede alterar el significado de los resultados; por ejemplo, la aplicación de una prueba en un lugar sin iluminación adecuada o muy ruidoso puede producir que varios sujetos no logren tener un buen desempeño en la prueba y llevarlos a no reflejar su verdadera habilidad, lo que implicaría una inadecuada interpretación sobre las puntuaciones y, por ende, los usos de las pruebas tendrían un bajo grado de validez.

Calificación de los ítems

En esta etapa se requiere de una guía de calificación. Los ítems cuyo formato de respuesta es cerrado son los más sencillos de calificar, ya que la guía solo debe contener las respuestas correctas. En el caso de las preguntas de respuesta abierta se requiere de una rúbrica o escala en la que: a) se esbocen las respuestas a esperar; y, b) se especifique la puntuación por cada una de las partes de las respuestas consideradas correctas (Mateo & Martínez, 2008). Además, las guías deben ser lo más exhaustivas posible con respecto a las distintas formas de resolución. Por otro lado, en la calificación de las preguntas de respuesta abierta se deben desarrollar procesos de equiparación por jueces para que las puntuaciones no estén sesgadas por las diferencias entre la severidad de los jueces. La variación en los criterios de calificación de los jueces es una amenaza a la validez, dado que lleva a que la puntuación considere elementos irrelevantes para la medición.

Una vez calificados los ítems se debe realizar el análisis estadístico de estos, con base en el modelo de medición establecido desde la primera etapa. En este sentido, se analizará la calidad psicométrica de las puntuaciones de cada ítem, por ejemplo discriminación y dificultad, y de la prueba en general según el modelo considerado. En caso de que existan ítems que no cumplan los estándares establecidos por la teoría, estos no deben ser considerados en la calificación final (Ferrara et al., 2017; Martínez-Arias et al., 2006; Muñiz & Fonseca-Pedrero, 2019).

Con el análisis de los ítems, se busca generar evidencias de validez basadas en la estructura interna por medio de la comprobación de la configuración factorial establecida en la teoría. También se buscan evidencias de precisión (o confiabilidad), es decir, que las puntuaciones de la prueba brinden una aproximación apropiada de los niveles de habilidad de los sujetos en el constructo medido por la prueba (AERA, APA, & NCME, 2014).

Generación de conclusiones

Una vez calificados los exámenes, se proce-

de a la generación de interpretaciones con respecto al análisis de las puntuaciones, y a determinar parámetros para las inferencias pretendidas. Si el objetivo es establecer niveles de dominio en las dimensiones de la prueba, se debe recurrir a un proceso de establecimiento de estándares con los ítems seleccionados (*standard setting*). Si la finalidad es comparar el rendimiento de los sujetos contra el resto de la población, se puede recurrir al análisis de los percentiles de las puntuaciones obtenidas. Por último, se debe entregar un informe de calificación a los sujetos, indicándoles el significado de estas, ya que una nota por sí misma no dice nada con respecto a su desempeño en la prueba y puede crear nociones erróneas a los examinados, dado que culturalmente algunas notas se asocian a un buen rendimiento mientras que otras a un bajo rendimiento.

Dadas las etapas teóricas para la construcción de una prueba educativa, a continuación se presenta la descripción de cada una de ellas en la elaboración de la Prueba de Habilidades Cuantitativas en el caso específico del test del 2018.

Etapas de la construcción de la PHC previas a las aplicaciones regulares

En el 2003, las autoridades de la UCR decidieron que era necesario crear una prueba de ingreso a las carreras cuyas mallas curriculares tenían varios cursos de matemática. La prueba que se construyó para este propósito fue la PHC, la cual se aplica regularmente todos los años, desde el 2015. La construcción de los formularios de cada año se basa en las etapas iniciales de construcción de pruebas (desde el plan general hasta el pilotaje), que se desarrollaron durante varios años de investigación. A continuación, se describen los elementos principales de cada una de esas etapas.

Plan general

En esta etapa se determinó que el constructo razonamiento cuantitativo (RC) es lo que se quiere medir con la PHC; mientras que el uso (*para qué*) establecido fue brindar un criterio para la selección de los nuevos estudiantes que quisieran ingresar a carreras que requerían del uso de la matemática. Cabe mencionar que el RC se define como "la habilidad para analizar información cuantitativa y determinar cuáles destrezas y procedimientos pueden ser aplicados para obtener la solución de un problema particular" (Dwyer, Gallagher, Levin, & Morley, 2003, p. 1). Esta definición indica que el RC está conformado por los procesos de razonamiento con contenidos matemáticos que se deben realizar para llegar a la solución de un problema específico. En consecuencia, para la evaluación del RC se debe garantizar que los contenidos sobre los que se desarrollan las tareas de RC sean conocidos por toda la población.

En línea con la definición del RC, se puede decir que este no es equivalente al conocimiento matemático, ya que en el primero no es importante qué cantidad de conocimiento matemático tenga el sujeto, sino que este pueda determinar cómo utilizar el conocimiento que domina en situaciones particulares. Entre los componentes más importantes que demanda el RC de acuerdo con Niss y Højgaard (2011) están:

- el pensamiento matemático: consiste en utilizar las propiedades de los objetos matemáticos para llegar a una conclusión;

- el abordaje de problemas: consiste en diseñar una estrategia para utilizar un concepto matemático en la resolución de un problema;

- el razonamiento matemático: consiste en determinar la validez de proposiciones matemáticas; y

- la representación: consiste en diseñar o in-

terpretar una representación necesaria para llegar a la solución de un problema.

Con respecto al uso establecido para la PHC, se sostiene que el RC es necesario para desenvolverse exitosamente en las profesiones que requieren de la matemática en su quehacer (Mayes, 2019; Ryan & Gass, 2017) dado que los especialistas en estas áreas deben determinar cómo utilizar la matemática para resolver un problema determinado. Contrario a los ejercicios matemáticos clásicos de las carreras universitarias, en muchas tareas laborales no se indica qué algoritmo matemático se debe emplear, sino que el profesional debe construir una estrategia eficiente y pertinente que permita el éxito en la tarea. Con este fin, los especialistas de diversas áreas deben atender interrogantes utilizando gran variedad de información cuantitativa; esto los obliga a tomar decisiones sobre cuáles estadísticos brindan la información más adecuada a la pregunta atendida. En otros casos, se debe comparar fenómenos modelados por medio de expresiones algebraicas, lo cual demanda que el profesional distinga cuál es la estrategia que le permite realizar la comparación solicitada. Es por esto que se considera que los estudiantes de los programas de estudio que utilizan la matemática en su campo laboral deben poseer un nivel aceptable de RC (Rojas, Mora, & Ordóñez, 2018).

Ahora bien, dado que con la PHC se pretende discriminar a los sujetos con habilidad aceptable de aquellos con habilidades menores, se estableció que se necesitaba generar un punto de corte que separara a estas dos poblaciones, el cual puede ser representado en una escala de habilidad normal estándar como el valor ($\theta = 0$). Esto implicó que se utilizara un modelo de interpretación de puntajes con base a criterios y que se utilizara a la TRI como modelo de medición, ya que esta teoría permite maximizar la precisión en un nivel de habilidad específico (Martínez-Arias et al., 2006; Muñiz, 1997).

Definición del contenido

Como se mencionó, los componentes del constructo RC establecidos en los ítems de la prueba son: el pensamiento matemático, el abordaje de problemas, el razonamiento matemático y la representación (Niss & Højgaard, 2011). Por otro lado, los conocimientos base que se utilizan en la prueba se estructuran de acuerdo con las áreas temáticas de la educación primaria y secundaria inicial de Costa Rica: análisis de datos, aritmética, álgebra y geometría. Solo se consideran contenidos de la secundaria inicial, dado que se asume que estos son los dominados por los estudiantes aspirantes a ingresar en las carreras mencionadas. La inclusión de contenidos poco dominados puede generar una fuente de varianza irrelevante al constructo pretendido.

Especificaciones del test

La tabla de especificaciones (Tabla 1) de la prueba se planteó como una matriz de procesos, según los componentes y áreas de contenido, con una distribución homogénea. Es importante mencionar que aunque se postulan distintos procesos, estos no se conciben como independientes entre sí (Niss & Højgaard, 2011), sino que el nombre del proceso refleja aquel que se considera más demandante en la solución del ítem. Por lo general los cuatro componentes se emplean para la resolución del reactivo en algún grado. Por lo tanto, se considera que el modelo factorial que mejor representa a los ítems es el unidimensional.

Por otro lado, como la población meta de la prueba era numerosa y el presupuesto limitado, se decidió recurrir a una prueba escrita de selección

única, con una hoja adicional en la que se consignen las respuestas. Es importante destacar que aunque la información brindada por los reactivos de selección única es valiosa, las preguntas de respuesta abierta podrían brindar mayor información de los procesos realizados por los sujetos, ya que se podrían evaluar los pasos de razonamiento efectuados para llegar a la respuesta o resolución del ítem.

Por otra parte, se estableció que los ítems debían tener una dificultad TRI promedio igual a 0, pues se buscaba maximizar la precisión en el nivel de habilidad ($\theta = 0$). Finalmente, por las características de los ítems, se determinó que 40 reactivos era una cantidad suficiente para alcanzar una precisión adecuada y que la resolución de una prueba de este tipo tomaría alrededor de dos horas. Cabe resaltar que se consideró que el cansancio asociado a resolver más ítems podría aumentar las fuentes de error de medición.

Tabla 1
Tabla de especificaciones pretendida para la PHC 2018.

Procesos	**Área de contenido**			
	Análisis de datos	**Aritmética**	**Álgebra**	**Geometría**
Pensamiento matemático	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems
Representación	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems
Abordaje de problemas	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems
Razonamiento matemático	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems	2 a 3 ítems

constructores entregaron los reactivos y estos fueron juzgados individualmente por el resto de los constructores y un juez adicional. En el juzgamiento se evaluó si los ítems realmente se ajustaban a los componentes de la tabla de especificaciones establecida y se analizó si tenían alguna de las fuentes de varianza irrelevante al constructo mencionadas previamente; luego, se procedió a un juzgamiento grupal para determinar si los ítems construidos podían ser considerados para el ensamblaje de la PHC.

Construcción de los ítems

El perfil para las personas encargadas de construir los ítems incluyó formación en enseñanza de la matemática y conocimientos de medición. En las capacitaciones de los constructores se presentaron el objeto de medida (RC) y sus componentes. Asimismo, se proporcionaron ejemplos de ítems en cada una de las categorías establecidas y se señalaron los elementos que agregaban varianza irrelevante al constructo: uso de algoritmos extensos, preguntas semejantes a las de contenido matemático, uso de tecnicismos y alta demanda de comprensión de lectura. Luego, a cada constructor se le solicitó que presentara un ítem de cada una de las casillas de la tabla de especificaciones.

Dos meses después de la capacitación, los

Pilotaje

Posteriormente, los ítems seleccionados fueron ensamblados en formularios con las características establecidas en la tabla de especificaciones y se aplicaron en muestras de estudiantes universitarios de primer año. Estas aplicaciones piloto permitieron desarrollar un banco de ítems de aproximadamente 80 reactivos con propiedades psicométricas adecuadas, según el modelo de TRI de dos parámetros. Estas propiedades se-

rán explicadas en la etapa de calificación de los ítems. Las puntuaciones de estas aplicaciones brindaron evidencias de validez basadas en la relación con otras variables (específicamente, con el rendimiento académico, Bolaños-Barquero & Rojas-Torres, 2013; Rojas et al., 2018; Rojas-Torres, 2014) y de estructura interna (Rojas, 2013).

Elaboración del formulario de la PHC 2018

La construcción del formulario final utilizado en el 2018 se realizó con 36 de los ítems incluidos en el banco y 4 ítems nuevos. El pilotaje de estos ítems nuevos se llevó a cabo en la aplicación real de la prueba. Lo anterior implicó que estos ítems no fueran tomados en cuenta en la calificación. Se decidió bajar la cantidad de ítems calificables para asegurar que los nuevos ítems tuvieran estadísticas basadas en la situación real de evaluación en vez de basarse en una aplicación piloto.

Ensamblaje de la prueba

La selección final de los ítems tuvo la distribución que se presenta en la Tabla 2. Entre los criterios utilizados para seleccionar a los ítems de la tabla de especificaciones se consideró que no debían existir ítems con procesos de resolución muy semejantes entre sí, ya que esto podría penalizar doblemente al sujeto que no logró determinar el proceso de resolución óptimo. También se descartaron los ítems cuyas dificultades TRI, registradas en las aplicaciones piloto, se alejaran de la dificultad promedio pretendida ($\theta = 0$). Para este formulario la dificultad TRI promedio fue de ($\theta = .123$).

Para la administración de la PHC se proporcionaron varias sedes de aplicación en distintos edificios de entidades educativas ubicadas en zonas estratégicas de Costa Rica. Cada una de estas sedes tuvo entre 2 y 11 aulas, según la cantidad de personas inscritas en cada sede. Además, se coordinó con las autoridades de las entidades para que durante la administración de la prueba no hubiese personas ajenas a la aplicación cerca de las aulas y para que las aulas tuviesen iluminación y escritorios en buen estado.

Para cada sede se conformó un equipo de aplicación compuesto por un coordinador general, aplicadores para cada aula y un asistente del coordinador. Cada coordinador fue capacitado por el equipo desarrollador de la PHC y, luego, los coordinadores de sede capacitaron a sus equipos. En la capacitación se indicaron las pautas que se debían seguir para la administración de la prueba, por ejemplo: la forma de ordenar los muebles (pupitres) del aula, el protocolo para permitir el ingreso de los examinados a las aulas, las consultas que se podían responder y la vigilancia del grupo.

El día de la aplicación de la prueba, se leyó a los estudiantes unas instrucciones generales de

Tabla 2
Tabla de especificaciones final de la PHC 2018.

Procesos	Área de contenido			
	Análisis de datos	Aritmética	Álgebra	Geometría
Pensamiento matemático	3 ítems	2 ítems	3 ítems	2 ítems
Representación	2 ítems	3 ítems	3 ítems	3 ítems
Abordaje de problemas	3 ítems	2 ítems	2 ítems	3 ítems
Razonamiento matemático	2 ítems	3 ítems	2 ítems	2 ítems

cómo resolverla. Algunas instrucciones se relacionaron con: los materiales que podían utilizar, el tiempo asignado para la resolución y la forma de rellenar la hoja para respuestas. El total de personas que tomó la prueba fue de 2387 (1112 hombres y 1275 mujeres; 787 estudiantes de colegio privado y 1600 de público). En su mayoría eran estudiantes de secundaria que deseaban ingresar a una de las carreras de la UCR que utilizaba la PHC para la selección de los nuevos estudiantes.

Calificación de los ítems

Dado el carácter empírico de esta etapa, se utilizará una estructura clásica de análisis estadístico (procedimiento-resultados) para explicar su implementación. Además, en esta etapa se presentará el método de calificación utilizado para procesos de investigación, el cual no ha sido implementado aún. Actualmente, se reportan las calificaciones con el número de ítems correctos.

Procedimiento. Primeramente, se analizó la calidad global del instrumento. Para esto se estimó el índice de confiabilidad de constructos de modelos de ecuaciones estructurales, el cual se considera satisfactorio si es superior a .70 (Cea-D'Ancona, 2002). Además, se evaluó la hipótesis de la unidimensionalidad mediante un análisis factorial confirmatorio (AFC), esta hipótesis representa la estructura interna teorizada para la prueba. En el AFC se consideró que un buen ajuste era alcanzado si la raíz del error cuadrático medio de aproximación (RMSEA) era menor que .06, el índice de Tucker-Lewis (TLI) era mayor que .95, el índice de ajuste comparativo (CFI) era mayor que .95 y las cargas factoriales de los ítems eran superiores a .30 (Hu & Bentler, 1999; Cea-D'Ancona, 2002).

Por otro lado, para evaluar la calidad de los ítems se estimó el modelo de TRI de dos parámetros. Se valoró que los ítems no tuvieran dificultades extremas (menores que -3 o mayores que 3). También, se analizó que su discriminación fuera mínimamente aceptable (mayor a .35) y que brindaran información en el punto de interés ($\theta = 0$; función de información mayor que .10; Martínez-Arias et al., 2006).

Luego, con los ítems que cumplieron todos los criterios, se estimó la habilidad de los sujetos. Seguidamente, se analizó el error estándar de la estimación para verificar si este era bajo en el punto que se pretendía establecer como punto de corte ($\theta = 0$). Finalmente, se concluyó cuántas personas estuvieron en los grupos de habilidad que se requería discriminar ($\theta > 0$ y $\theta < 0$).

Análisis estadístico. Todos los análisis estadísticos se realizaron en la plataforma de programación estadística R, en su versión 3.3.2 (R Core Team, 2016). Los paquetes utilizados fueron *lavaan* (Rosseel, 2012), para la estimación de AFC, y *mirt* (Chalmers, 2012), para la estimación del modelo TRI.

Resultados. El índice de confiabilidad de constructos fue de .91, por lo cual se concluye que la prueba muestra evidencias de confiabilidad. Esta evidencia se basa en que los ítems son consistentes en la evaluación del constructo, es decir, las asociaciones de las puntuaciones de los ítems entre sí son semejantes. El indicador de consistencia interna usual de la TCT es el alfa de Cronbach, el cual fue de .85.

La estimación del modelo unidimensional del AFC se ajustó aceptablemente a los ítems de banco utilizados (CFI = .992, TLI = .992 y RMSEA = .014). No obstante, hubo 3 ítems (28, 29 y 35) que presentaron cargas factoriales inferiores a .30. El ajuste de este modelo brindó una *evidencia de validez basada en la estructura interna*.

Posteriormente, se analizaron las propieda-

des de los ítems según el modelo de TRI de dos parámetros. Se obtuvo que los mismos tres ítems que presentaron cargas factoriales bajas presentaban problemas en la discriminación (lo cual es esperable por la analogía de la carga factorial de un modelo unidimensional con la discriminación de la TRI de dos parámetros) y en la información en $\theta = 0$ (no solo fueron menores a .10, sino que fueron menores a .03). Hubo otros dos ítems (9 y 13) que presentaron valores de información en $\theta = 0$ ligeramente menores que el umbral establecido (.09 y .08), pero cumplieron el resto de los criterios establecidos. Por tanto, se decidió eliminar únicamente a los primeros tres ítems para la estimación de la habilidad.

Finalmente, la *calificación* de los sujetos correspondería a la habilidad de los sujetos a partir de los ítems seleccionados. Se obtuvo que el 44.0% de la población (1311 sujetos) tenía una habilidad mayor o igual a $\theta = 0$. Por otro lado, se recurrió a una de las ventajas de la TRI: la estimación del error estándar de la habilidad, lo cual permite la generación de intervalos de confianza. Se realizó la prueba de hipótesis para la hipótesis nula de $\theta \geq 0$, esta se rechazó con una significancia del 5% para el 30.5% de la población (911 personas); es decir, que el 30.5% de los sujetos presentaron una habilidad inferior a 0, con un 95% de confianza. De forma análoga, se obtuvo que el 24.7% de la población (739 personas) tuvo una habilidad superior a 0, con una confianza del 95%.

Generación de conclusiones

A partir de los datos de la calificación, la generación de conclusiones con propósitos de investigación determinaría que con un 95% de confianza, el 30.5% de la población no tiene la habilidad de RC requerida y que el 24.7% sí posee la habilidad requerida, por tanto, los primeros no pueden concursar por aquellas carreras en la que se utiliza la PHC; mientras que los segundos sí pueden participar. Con respecto al 44.7% de la población restante, no se pueden generar conclusiones tan contundentes.

La generación de conclusiones implementada oficialmente por la mayoría de las carreras se basa en un umbral asociado al porcentaje de respuestas correctas. Se considera que los sujetos con notas menores a este umbral no poseen la habilidad requerida para cursar estas carreras. El umbral se estableció con el cuidado de que las personas con notas inferiores a este punto presentaran habilidades inferiores a 0. A los sujetos con notas debajo del umbral se les indicó que no cumplían con el requisito para concursar en las carreras. La razón por la que no se usa la habilidad TRI como criterio oficial es que es difícil de informar a los sujetos, ya que dos sujetos con la misma cantidad de aciertos no tendrán necesariamente el mismo nivel de habilidad.

Discusión

Este artículo muestra que el proceso de construcción de una prueba educativa es complejo y demandante. Es un proceso que no se puede realizar de manera ligera e irreflexiva. No obstante, la masificación del uso de estos instrumentos de evaluación ha generado que muchas de las etapas sean ignoradas. Esto induce a que las inferencias y los usos de las puntuaciones de las pruebas carezcan de validez.

En el caso de la PHC, las etapas iniciales de plan general, definición del contenido y especificaciones del test *requirieron* un período de reflexión de aproximadamente dos años; en esta discusión estuvieron involucrados expertos en medición, representantes de las carreras intere-

sadas y autoridades universitarias. El período de tiempo puede parecer excesivo, pero cuando se considera el uso propuesto para las puntuaciones de la prueba este tiende a ser razonable. En el caso de las pruebas educativas utilizadas en los salones de clase, se requiere que el tiempo dedicado a estas etapas iniciales sea mucho menor, pero esto no es excusa para eliminar la reflexión sobre ¿qué se desea medir?, ¿para qué se desea medir?, y ¿cómo se debe medir? Estas reflexiones pueden lograr evitar errores de medición clásicos de evaluación educativa.

En las etapas de construcción de los ítems y ensamblaje de la prueba se debe procurar que el formulario a desarrollar responda a las preguntas planteadas previamente. Además, en estas etapas, junto con la administración de la prueba, se debe procurar que los examinados tengan las condiciones necesarias para que logren mostrar su verdadero nivel de habilidad. En el caso de la PHC se invierte un período de aproximadamente ocho meses para el diseño de un formulario, este ciclo comienza con la construcción de los nuevos ítems, luego, el juzgamiento de los ítems, el ensamblaje de la prueba y la revisión del contenido del formulario. Como ya se mencionó, para las pruebas de aula no se puede exigir estos periodos de tiempo, pero no pueden ser eximidas de las reflexiones sobre si los ítems utilizados y la prueba en su totalidad miden lo que se desea medir y si las condiciones de aplicación realmente posibilitan que el examinado muestre su verdadera habilidad.

Con respecto a la calificación se debe tener la flexibilidad de variar la puntuación en función de la calidad de los ítems, ya que los mecanismos rígidos de calificación pueden llevar a interpretaciones erróneas, lo cual repercute drásticamente en la generación de conclusiones, y por ende perjudica a los examinados. Al menos, se debe procurar brindar a los sujetos las conclusiones reales relacionadas con su rendimiento, dado que muchas veces hay creencias generalizadas sobre los significados de las calificaciones que no corresponden con el planteamiento teórico de la prueba; por ejemplo, un buen rendimiento es superior a un 7 de la escala de 0 a 10. En el caso de la PHC, actualmente se está trabajando en crear un informe de resultados que sea lo suficientemente informativo.

La última etapa en la construcción de una prueba es una de las más fallidas, pocas veces se le dice a un examinado qué significa la puntuación que obtuvo. Esta omisión es un problema ético importante, dado que se puede afectar la autoestima de un sujeto por una interpretación errónea, la cual es esperable debido a la atmósfera cultural. Debido a esto, en los estándares de evaluación psicológica y educativa se enfatiza sobre la forma de entregar los informes de calificación (AERA, APA, & NCME, 2014). Ahora bien, la interpretación de los puntajes solo es posible si la construcción de la prueba fue realizada siguiendo detalladamente las etapas de construcción de la misma, por tanto, se hace evidente la importancia de los protocolos de construcción de pruebas.

Es importante mencionar que la construcción de una prueba es un proceso continuo. En el caso de la PHC, después de los resultados de la aplicación del 2018, se concluyó que es necesario desarrollar un ensamblaje con ítems cuyas dificultades fueran cercanas a 0, en vez de que su promedio fuera cercano a 0, dado que esto permitiría aumentar la discriminación en el punto deseado y de esta manera disminuir el porcentaje de sujetos que no se puede clasificar en habilidades mayores que 0 o menores que 0. El abandono del promedio de la dificultad TRI se debe a que los ítems maximizan su discriminación en el nivel de dificultad; el promedio no implica ítems con dificultades en el nivel deseado. Por otro lado, el equipo constructor de la PHC está trabajando en

una definición de procesos más específicos para ciertos ítems de RC, en vez de los mencionados en este artículo, que son más comunes a todos los ítems de la prueba (Jeannotte & Kieran, 2017).

Finalmente, es importante mencionar que este artículo presenta un planteamiento general de las etapas necesarias para la construcción de una prueba educativa y que utiliza como ejemplo una prueba escrita de selección única, con un modelo de interpretación de puntajes basado en criterios y con modelo de medición de la TRI de dos parámetros. Los detalles específicos para cada variante de una prueba educativa no pueden ser presentados en un artículo de revista, ya que dependerán de la finalidad de la medición. No obstante, las etapas presentadas en este artículo incluyen los elementos mínimos que debe considerar cualquier prueba educativa.

Referencias

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington D. C.: American Educational Research Association.

Bolaños-Barquero, M., & Rojas-Torres, L. (2013). Comparación entre los promedios de la Prueba de Aptitud Académica y la Prueba de Habilidades Cuantitativas de los estudiantes de la universidad de Costa Rica. *Revista de Ciencias Sociales, 142*(IV), 101-115. doi: 10.15517/rcs.v0i142.14305

Castillo-Arredondo, S., & Cabrerizo-Diago, J. (2010). *Evaluación educativa de aprendizajes y competencias.* Madrid, España: Pearson Educación.

Cea-D'Ancona, M. A. (2002). *Análisis multivariable. Teoría y práctica en la investigación social.* Madrid, España: Síntesis.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal Of Statistical Software, 48*(6), 1-29. doi: 10.18637/jss.v048.i06

Downing, S. M. (2006). Twelve steps for effective test development. En S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 3-26). Londres, Inglaterra: Lawrence Erlbaum. doi: 10.4324/9780203874776.ch1

Dwyer, C. A., Gallagher, A., Levin, J., & Morley, M. E. (2003). What is quantitative reasoning? Defining the construct for assessment purposes. *ETS Research Report Series, 2003*(2), 1-48. doi: 10.1002/j.2333-8504.2003.tb01922.x

Embretson, S. (2017). An integrative framework for construct validity. En A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications* (pp. 102-123). Oxford, MS: Willey Blackwell. doi: 10.1002/9781118956588

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled approaches to assessment design, development and implementation. En A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications* (pp. 41-74). Oxford, MS: Willey Blackwell. doi: 10.1002/9781118956588

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Jeannotte, D., & Kieran, C. (2017). A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in Mathematics, 96*(1), 1-16. doi: 10.1007/s10649-017-9761-8

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi: 10.1111/jedm.12000

Martínez-Arias, M. R., Hernández-Lloreda, M. J., & Hernández-Lloreda, M. V. (2006). *Psicometría.* Madrid, España: Alianza.

Mateo, J., & Martínez, F. (2008). *Medición y evaluación*

educativa. Madrid, España: La Muralla.

Mayes, R. (2019). Quantitative reasoning and its rôle in interdisciplinarity. En B. Doig, J. Williams, D. Swanson, R. Borromeo-Ferri & P. Drake (Eds.), *Interdisciplinary Mathematics Education. The State of the Art and Beyond ICEM 13 Monographs* (pp. 113-133). Cham, Suiza: Springer. doi: 10.1007/978-3-030-11066-6_8

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. doi: 10.3102/0013189X018002005

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems.* Madrid, España: Pirámide.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema, 31*(1), 7-16. Recuperado de http://www.psicothema.com

Niss, M., & Højgaard, T. (2011). *Competencies and Mathematical Learning. Ideas and inspiration for the development of mathematics teaching and learning in Denmark.* Roskilde, Dinamarca: IMFUFA.

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Recuperado de https://www.R-project.org

Rojas, L. (2013). Predicción de la dificultad de la Prueba de Habilidades Cuantitativas de la Universidad de Costa Rica. *Revista Digital Matemática, Educación e Internet, 13*(1), 1-14. Recuperado de https://tecdigital.tec.ac.cr/revistamatematica

Rojas-Torres, L. (2014). Predicción de la reprobación de cursos de matemática básicos en las carreras de Física, Meteorología, Matemática, Ciencias Actuariales y Farmacia. *Revista Electrónica EDUCARE, 18*(3), 3-15. doi: 10.15359/ree.18-3.1

Rojas, L., Mora, M., & Ordóñez, G. (2018). Asociación del razonamiento cuantitativo con el rendimiento académico en cursos introductorios de matemática de carreras STEM. *Revista Digital Matemática, Educación e Internet, 19*(1), 1-13. Recuperado de https://tecdigital.tec.ac.cr/revistamatematica

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal for Statistical Software, 48*(2), 1-36. doi: 10.18637/jss.v048.i02

Ryan, A. M., & Gass, S. E. (2017). Quantitative reasoning: Exploring troublesome thresholds'. *Discussions on University Science Teaching: Proceedings of the Western Conference on Science Education, 1*(1), 1-16. Recuperado de https://ir.lib.uwo.ca/wcsedust

Tiana, A. (1996). La evaluación de los sistemas educativos. *Revista Iberoamericana de Educación, 10*, 37-61. Recuperado de https://rieoei.org/RIE

Villareal-Galera, M. P., Alfaro-Rojas, L., & Brizuela-Rodríguez, A. (2015). *Construcción de pruebas estandarizadas en el ámbito de la medición educativa y psicológica. Serie Cuadernos Metodológicos del IIP*. San José, Costa Rica: Instituto de Investigaciones Psicológicas. Recuperado de http://www.kerwa.ucr.ac.cr