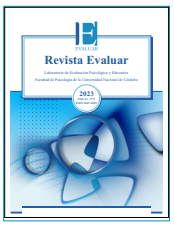
2023, Vol. 23, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

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



Elementos metodológicos de thinking aloud para la obtención de evidencias de validez de contenido

Thinking aloud methodological elements to obtain evidence of content validity

Graciela Ordóñez-Gutiérrez * [1]

1 - Universidad de Costa Rica.

Recibido: 04/07/2023 **Revisado:** 20/09/2023 **Aceptado:** 27/09/2023

Resumen

Los test *thinking aloud* se han implementado en la investigación educativa, principalmente, para obtener las estrategias de solución de los sujetos en diferentes tareas. En particular, se han empleado con la finalidad de determinar las estrategias que los examinados implementan para dar solución a ítems en pruebas educativas. Sin embargo, los elementos metodológicos para elaborar un *thinking aloud* paso a paso y obtener evidencias de validez no son explícitos en la literatura. Sin embargo, es importante obtener esos resultados, por lo que el objetivo de este artículo es proporcionar los pasos metodológicos, desde una perspectiva práctica de implementación en la obtención de habilidades de razonamiento cuantitativo. Para llevar a cabo los test *thinking aloud*, se realizó una revisión exhaustiva de la literatura y se propusieron las siguientes etapas: 1) definición del propósito, 2) elaboración de los procesos de solución de los ítems desde una perspectiva teórica y práctica, 3) selección de la muestra, 4) proceso de simulación, 5) recolección de los datos, 6) transcripción y análisis. Al seguir estas etapas, se pueden efectuar los test *thinking aloud* que resulten exitosos y que permitan obtener evidencias válidas y confiables.

Palabras clave: *thinking aloud, evidencias de validez, selección de la muestra, índice de consistencia de expertos, habilidades de razonamiento cuantitativas*

Abstract

Think-aloud has been implemented in educational research mainly to define the solution strategies of the subjects on different tasks. In particular, it has been used to determine the strategies the subjects implement to find solutions to items in educational tests. However, since the steps of the methodological think-aloud elements followed to obtain valid evidence are not explicit and are important to define, this article aims to provide the methodological steps, from a theoretical perspective so they can be implemented. To develop the steps, a comprehensive literature review was carried out, and the following steps were proposed: 1) define the purpose, 2) elaborate the solution processes of the items from a theoretical perspective, 3) select the thinking-aloud sample, 4) simulation process, 5) data recollection, 6) think-aloud transcript and analysis. These steps can lead to a successful think-aloud process that produces reliable, valid evidence.

Keywords: *thinking aloud, valid evidence, select sample, expert consistency index, quantitative reasoning skill*

***Correspondencia a:** graciela.ordonez@ucr.ac.cr
Cómo citar este artículo: Ordóñez-Gutiérrez. G. (2023): Elementos metodológicos de *thinking aloud* para la obtención de evidencias de validez de contenido. *Revista Evaluar*, *23*(3), 45-60. Recuperado de https://revistas.unc.edu.ar/index.php/revaluar
Participaron en la edición de este artículo: Vanesa Mariela Toledo, Stefano Macri, Juan Cruz Balverdi Nieto, Pablo Carpintero, Florencia Ruiz, Rodrigo Maderna, Jorge Bruera.

Introducción

En los *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) para la elaboración de pruebas educativas y psicológicas, se establece que, para determinar la calidad de medida de los test, es indispensable recolectar evidencias de validez en las que se justifiquen las inferencias que se realizan sobre las puntuaciones que obtiene un grupo de personas. Una de las evidencias requeridas para realizar inferencias válidas sobre las puntuaciones de dichos test hace referencia a la recolección de los procesos de respuestas, o bien estrategias que emplean los sujetos para dar solución a los ítems que componen la prueba (Green, 1998; Embretson, 2017; Keehner et al., 2017; Padilla & Benítez, 2014; Leighton, 2013) y que forman parte de las evidencias relativas a la validez del constructo.

Con la recolección y análisis de las estrategias, se obtendrían evidencias de validez de contenido (Padilla & Leighton, 2017; Fonteyn et al., 1993) y al mismo tiempo permitiría verificar si la prueba realmente mide el constructo que se pretende medir. Además, con los test thinking aloud, también llamados en la literatura como "protocolos de pensamiento en voz alta", "protocolos verbales" e incluso "entrevistas cognitivas", se pueden detectar posibles fuentes de varianza irrelevante al constructo (Brizuela et al., 2016) y el Funcionamiento Diferencial de los Ítems (DIF, por sus siglas en inglés) (Ercikan et al., 2010). Por lo tanto, los test *thinking aloud* se han empleado para determinar habilidades de razonamiento de las personas y juegan un papel importante en la obtención de evidencias de validez de instrumentos de medición, así como en la elaboración de metodologías de evaluación (Green, 1998). En este sentido, y según Keehner et al. (2017), los test thinking aloud aplicados a las tareas de evaluación brindan una perspectiva única sobre el procesamiento individual de los sujetos incluyendo la información referente a los conceptos erróneos, las debilidades en las habilidades y el uso de estrategias para la resolución de problemas.

En otro sentido, Padilla y Leighton (2017) manifiestan que la falta de experiencia de los investigadores, la consolidación de las mejores prácticas para efectuar los test *thinking aloud* y las recomendaciones de cómo obtener evidencia de los procesos de respuesta pueden generar que se pierda la oportunidad de recolectar datos valiosos y sustantivos y resultar en evidencias que lleven a interpretaciones incorrectas sobre las puntuaciones de los examinados. Por otro lado, a pesar de la importancia y de la existencia de una amplia literatura sobre las maneras de efectuar reportes verbales, estos no brindan elementos explícitos y que son indispensables a la hora de realizar los test *thinking aloud* desde la perspectiva de las pruebas educativas para obtener evidencias de validez. Por ello, el objetivo de este artículo es proporcionar las etapas metodológicas, desde una perspectiva teórica y práctica, para efectuar los test *thinking aloud* y obtener evidencias de validez de contenido.

Método

Para elaborar la perspectiva teórica se realizó una búsqueda en sitios web. Primeramente, se escribió en el sitio *scholar.google.com* la expresión "*thinking aloud in education testing*", la cual arrojó 41 mil resultados en julio del 2019. Luego, se realizó la búsqueda empleando la frase "*thinking aloud in education items testing*" y la cantidad de resultados se redujo a 12. Por otra parte, también se buscó la frase "*verbal protocols*

in education testing" y esto llevó a 13 resultados. Asimismo, se efectuó una búsqueda en las bases de datos *ProQuest* y *Erick*.

Es importante mencionar que, la mayoría de estos escritos tienen como contexto el deportivo, la neurociencia, la psicología, la enfermería, la neuropsicología, entre otras, que están más asociadas a la Ciencias Médicas que a la educación (Ercikan et al., 2010). Por esta razón, la cantidad de escritos que consideran los test *thinking aloud* para la recolección de evidencias desde la perspectiva educativa en general, más precisamente en pruebas educativas, es escasa.

Ahora bien, los documentos que hacen referencia a alguna prueba educativa están enfocados en los procesos de solución de ítems en el área de idiomas (*language*) y área verbal (en el caso del español). Cabe resaltar que, los artículos encontrados se analizaron con detalle y de estos, se extrajo y se compararon las etapas que los investigadores siguieron para realizar los test *thinking aloud*, lo que dio como resultado una cantidad de 6 etapas para su elaboración.

Para elaborar la perspectiva práctica, se realizaron los test *thinking aloud* a 13 personas (7 mujeres y 6 hombres), con el objetivo de determinar habilidades de razonamiento cuantitativo mediante la aplicación de un test. Para esta aplicación, se utilizó el formulario de la Prueba de Habilidades Cuantitativas, la cual está compuesta por 40 ítems de selección única con cuatro opciones de respuesta. Cada aplicación se realizó bajo estricta confidencialidad y de manera individual. Además, se aplicó de manera concurrente y después retrospectiva.

Etapas teóricas de los test thinking aloud

El test *thinking aloud* es un método empleado en estudios asociados a la psicología cognitiva para recolectar información acerca de los procesos de información humana en la resolución de un problema que requiere de la manipulación de información para encontrar la solución de un estado de cosas complejo (Padilla & Leighton, 2017; Fonteyn et al., 1993). Actualmente, a este método lo emplean algunos investigadores para explorar la validez de los ítems en pruebas de rendimiento (Leighton, 2004; Padilla & Leighton, 2017; Embretson, 2017), pues proporciona información sobre los procesos cognitivos generados por los examinados, y por ende, para determinar evidencias de validez que permitan inferencias sobre las puntuaciones del test. Además, permite fortalecer el constructo que se pretende medir con la prueba. Por ello, la manera de llevar a cabo los test *thinking aloud* permitirá recolectar la información suficiente y necesaria de forma rigurosa. Por este motivo, a continuación, se explican las etapas que permitirán la elaboración de reportes verbales exitosos en el contexto de pruebas educativas.

Es importante mencionar que, existe una coincidencia entre la literatura sobre la preocupación de la confiabilidad de los datos obtenidos con los test *thinking aloud*, ya que las verbalizaciones de los sujetos pueden reflejar lo que un grupo de examinados piensa, en lo que cree que están haciendo y no sus procesos de pensamiento reales (Ercikan et al., 2010). Ante esto, en Leighton (2004) se exponen algunos problemas relacionados con la confiabilidad de las verbalizaciones y, también, se pueden encontrar recomendaciones al respecto. Específicamente, Leighton (2004) hace referencia a tres factores que afectan el uso exitoso de los test *thinking aloud*: 1) el momento en que se recolectan los informes, 2) la manera en que se recolectan, y 3) las tareas, o bien los ítems, seleccionados para la recolección de los procesos. Por eso, se recomienda seguir una detallada y rigurosa metodología en la que se resguarde la confiabilidad de los reportes. A continuación, se

detallan los pasos sugeridos en la literatura para realizar los test *thinking aloud*.

Definir el propósito de efectuar los test thinking aloud

Esta primera etapa es indispensable y no se encuentra explícita en la literatura revisada, por lo que su planteamiento se establece a partir de la experiencia de la autora de este trabajo. Pues, definir *qué* y *para qué* realizar los test *thinking aloud* permite al investigador tener un panorama claro de lo que realmente se requiere y se pretende con ello. Además, dicha definición le genera los indicadores que necesitará en el momento de implementar los test thinking aloud con las personas examinadas. Luego, se deberá elaborar una guía con los elementos que se incluyeron en la definición del *qué* y *para qué*.

Por otra parte, se deberán elegir los ítems de la prueba que presenten las mejores medidas métricas según alguna teoría de medición (esta puede ser a partir de la Teoría Clásica de los Test, Teoría de Respuesta a los Ítems, u otra), según la dificultad y la discriminación. Está claro que, esto depende de la finalidad de efectuar los test *thinking aloud* en la obtención de evidencias de validez del constructo, ya sea que con el constructo se quieran medir habilidades o conocimientos. Si la finalidad es recolectar información con respecto a las dificultades o errores que comenten los examinados, entonces, se deberá elaborar un listado de los posibles errores para, luego, determinar qué elementos son considerados por el grupo de personas a la hora de cometerlos. Si la finalidad es determinar las habilidades, entonces, se deberá contar con una serie de indicadores que guíen la identificación de esas habilidades.

Elaboración de los procesos de cada ítem

Luego de seleccionar los ítems, la persona investigadora deberá considerar jueces expertos en el tema de interés, para que resuelva cada uno de los reactivos. De ser posible, cada ítem deberá estar enmarcado en alguna dimensión, según el constructo a medir, o bien, según la finalidad de medición mediante la prueba. En esta etapa, cada juez debe valorar si el reactivo pertenece o no a la dimensión del constructo seleccionado y resolverlo proporcionando la solución, según sus requerimientos. Es importante aclarar que, los jueces deben ser expertos en el área de interés.

Después de elaborar los procesos dados por los jueces, la persona investigadora deberá realizar el análisis de congruencia sobre los procesos generados por los expertos y verificar que todos siguen el mismo proceso de solución, o bien, llegar a un consenso entre ellos para establecer el mejor camino de solución y que responda a las dimensiones del constructo. Dicho análisis se podrá realizar empleando algún índice de congruencia entre jueces, como por ejemplo, Kappa de Cohen (Landis & Koch, 1977); Kappa de Fleiss (Fleiss, 1971; Cohen, 1960), el índice de Rovinelli y Hambleton (1977), *Many-Facet-Rasch Measurement* (Prieto-Adánez, 2011), el índice de consistencia de *Hierarchy* (Leighton et al., 2009), la metodología *Rule-Space* (Artavia-Medrano, 2015) o la Teoría de la Generalizabilidad (Martínez-Arias et al., 2006).

Elección del tamaño de la muestra

La elección del tamaño de la muestra, o de la cantidad de personas que se elegirán para efectuar los test thinking aloud es indispensable, ya que con ello se determinará la cantidad de datos que se producirán (Padilla & Leighton, 2017; Leighton,

2004). Además, dependiendo de la personalidad de cada participante, se pueden obtener datos fructuosos en información, o por lo contrario, se pueden obtener datos bajos en información. Desde esta perspectiva, Fonteyn et al. (1993) argumenta que, lo mejor es recolectar datos ricos y profundos de una pequeña muestra de sujetos.

Con respecto a la muestra de sujetos, varios estudios (Fonteyn et al., 1993; Virzi, 1992) sugieren que una cantidad de 5 o 6 personas puede producir resultados estables. No obstante, otros investigadores indican que una muestra pequeña y heterogénea de sujetos afecta la relación entre el tamaño de los datos y su estabilidad (Van Den Haak et al., 2003). Cabe aclarar que el tamaño de la muestra dependerá de la finalidad de realizar los test *thinking aloud*. En este sentido, Padilla y Benítez (2014) manifiestan que, para determinar el número total de los participantes, se deben tener dos criterios: 1) la saturación teórica y 2) la relevancia de la información. Ambos criterios afectan el muestreo porque el número y las características de los sujetos dependen del análisis de las entrevistas. De esta manera, Padilla y Benítez (2014) indican que la saturación teórica hace referencia a que los investigadores deberán seguir aplicando los test *thinking aloud* hasta que no surjan nuevos hallazgos, mientras que la relevancia teórica hace referencia a la selección de los entrevistados, de acuerdo con la teoría del constructo de la investigación.

Con respecto al primer criterio, lo que se pretende es construir teoría, mientras que con el segundo se pretende comprobar una teoría. Ahora bien, si se quiere construir una teoría, entonces, la cantidad de sujetos a considerar en el muestreo es entre 20 y 50, pero si se requiere comprobar la teoría, basta con considerar 10 sujetos, ya que los patrones de respuestas en función de la teoría se vuelven repetitivas a partir del séptimo sujeto (Padilla & Benítez, 2014). Es importante mencionar que, estos criterios deben ser considerados en términos del objetivo del estudio y de la complejidad de la evaluación propuesta.

Entrenamiento de las personas participantes

Antes de realizar los test *thinking aloud* cuyos datos formarán parte del estudio, es importante entrenar a las personas participantes del estudio; esto permitirá al grupo de sujetos tener una visualización de lo que se quiere que se realice en el estudio. Además, se puede anticipar qué dificultades tendrá el participante en la verbalización de los procesos de respuesta, y así, poder controlarla en la recolección de los datos reales (Padilla & Leighton, 2017). En este sentido, Fonteyn et al. (1993) manifiestan que, el investigador puede hacer una valoración sobre la relevancia de las verbalizaciones de los sujetos, de la misma tarea a proporcionar, de la preselección de la tarea, o bien, de la escogencia del ítem, de la estandarización en cuanto a la forma de llevar a cabo el trato con el sujeto y, además, una valoración de aspectos emergentes a la hora de realizar los test *thinking aloud* necesarios para su investigación. Además, anticipará formas adicionales de abordar un mismo tema y de controlar el tiempo real con la finalidad de estandarizarlos para todos los examinados (Fonteyn et al., 1993). Ahora bien, es crucial tomar en cuenta que, aunque los entrenamientos permiten establecer una presentación de un conjunto estándar de circunstancias, estas no son iguales para todo el conjunto de participantes, pues son diferentes a las que se pueden presentar en las condiciones reales. Sin embargo, sí proporcionan una idea inicial de las diferentes circunstancias que se puedan presentar.

Recolección de los datos

Padilla y Leighton (2017) y Fonteyn et al. (1993) argumentan que, cada persona participante tiene que ser convocado en un horario en el cual la persona investigadora y participante puedan. Las sesiones deben darse de manera individual y el lugar debe ser un entorno tranquilo que facilite el test *thinking aloud* de los sujetos y en las que no se generen distracciones.

Además, todas las sesiones de los test *thinking aloud* deben ser grabadas en audio, seguida de la transcripción de los datos verbales producidos por los entrevistados. Por otra parte, se les debe decir a las personas participantes que piensen constantemente en voz alta, sin dejar que se generen pausas prolongadas (entre dos o más minutos), ya que con esto se perdería información sobre los procesos de pensamiento. Si las personas participantes realizan una pausa un poco prolongada, entonces el investigador debe recordarles que deben resolver los ítems o la tarea expresando en voz alta la manera en que van resolviéndola. No obstante, se debe recordar que la interacción entre el investigador y el sujeto debe ser mínima mientras se estén efectuando los test *thinking aloud* con mínimas interrupciones.

En Green (1998), se indican algunas recomendaciones para la recolección de los datos antes de iniciar las sesiones de los test *thinking aloud*. Algunas de estas son:

1 - Preparar instrucciones sin ambigüedades en las que se explicite lo que se debe hacer con el objeto, la tarea o el ítem; es decir, una instrucción clara.

2 - Informar a las personas sobre lo que realmente se requiere de ellos y explicar claramente cuál es el procedimiento y las condiciones de este.

3 - Practicar con la persona al menos un par de veces para que adquiera confianza y no se intimide ante la tarea que debe ejecutar.

4 - Aclarar a la persona participante que debe mantenerse hablando; si se mantiene en silencio por periodos prolongados, realice la siguiente pregunta: ¿puedes indicarme qué estás pensando?

5 - Durante la sesión, estar atento a los silencios prolongados de las personas participantes.

6 - Permitir cierto tiempo adicional si no se ha terminado la tarea, puesto que la verbalización requiere de más tiempo que el simple hecho de escribir y elegir una respuesta correcta.

7 - Utilizar un equipo adicional como videocámaras, o grabaciones en audio, pues no se puede recolectar todo con solo escuchar y escribir en un reporte.

8 - Finalmente, después de las sesiones, efectúe preguntas que le permitan corroborar lo expresado por la persona participante a la hora de realizar la tarea que se indicó. En este caso, se le puede pedir que explique el proceso realizado para resolver la tarea.

Desde otra perspectiva, Fonteyn et al. (1993), recomiendan que las verbalizaciones de los participantes se deben dividir en dos partes: una concurrente y otra retrospectiva. Si la información se verbaliza en el momento en que el sujeto resuelve la tarea o el ítem, esta será una verbalización concurrente. Pero, si se le indica a la persona participante que explique el proceso llevado a cabo para los ítems, o la tarea, luego de haberlos resuelto, esta será una verbalización retrospectiva (Ericsson & Simon, 1987; Russo, Johnson & Stephens, 1989; Sapsirin, 2016). Ambos tipos de verbalización son importantes para determinar evidencias de validez en pruebas educativas y también de instrumentos de medición. En este sentido, la información que se obtenga del proceso retrospectivo complementa los datos concurrentes.

Transcripción, codificación y análisis de los thinking aloud

La transcripción de los datos es la parte más larga y lenta de efectuar. Sobre todo, si no se cuenta con los softwares modernos que permitan la transcripción de los datos de audio a datos escritos. Sin embargo, es una de las etapas que se deben de realizar para efectuar la codificación y, posteriormente, el análisis de los datos. La transcripción y codificación de los test *thinking aloud* deben ser un proceso riguroso y estandarizado, que incluya múltiples evaluadores y el cálculo de un índice de confiabilidad, como los mencionados en el apartado de elaboración de los procesos de solución de los ítems (Padilla & Leighton, 2017; Ericsson & Simon, 1987; Green, 1998).

La persona investigadora debe realizar las transcripciones de manera individualizada; esto es, por cada sujeto participante, de forma tal que se respete el lenguaje y las terminologías empleadas por cada uno. Luego de la transcripción, las verbalizaciones de los participantes se pueden dividir en secciones o en líneas de unidades o segmentos etiquetados, de acuerdo con la teoría establecida previamente, o según la finalidad de la prueba. Estas divisiones facilitan la identificación de partes del texto que son importantes para determinar puntos o palabras claves de la investigación (Joseph & Patel, 1990)

Ahora bien, durante la etapa de transcripción, es importante tener en cuenta: 1) la creación de una codificación de los test *thinking aloud* basada en un modelo teórico del constructo a medir sobre la tarea o proceso desarrollado por las personas participantes, 2) la inclusión del conjunto de conocimientos y habilidades esperadas, y ejemplos de los tipos de respuestas que se esperaría como evidencia de los conocimientos y habilidades; 3) si fuera el caso, la capacitación a los evaluadores que efectuarán la revisión y valoración de los procesos de solución de cada uno de los ítems realizados por los examinados en los test *thinking aloud* y los esperados teóricamente; 4) en el entrenamiento de los evaluadores, la indicación de que no valoren la dificultad del ítem, la discriminación, ni tampoco el funcionamiento diferencial de los ítems, sino que se centren en los procesos de respuesta brindado por cada uno de los participantes; 5) el cálculo del acuerdo inicial entre los evaluadores y determinación de la fuerza de concordancia entre las puntuaciones otorgadas en los procesos realizados versus los procesos esperados; 6) si la fuerza de concordancia entre los jueces es baja, entonces se debe volver a realizar una valoración con otros jueces, con la finalidad de obtener una puntuación de concordancia aceptable (mayor a 0.60 en Kappa por ejemplo).

En otro sentido, Ericsson y Simon (1984) manifiestan que, el análisis de los protocolos conlleva tres pasos: 1) un análisis de las frases de referencia; 2) un análisis afirmativo; 3) un análisis de guiones. En el primer paso, el investigador debe identificar todos los nombres y frases nominales en los datos verbales de cada sujeto y codificar las frases con el nombre del concepto de referencia. Estos códigos definen e identifican el vocabulario que los sujetos han verbalizado durante la ejecución de la tarea. Luego, cada concepto es identificado y codificado, de acuerdo con la identificación del investigador y según la naturaleza de la tarea. Ahora bien, el análisis de las frases de referencia continúa, según Fonteyn et al. (1993), hasta que todos los conceptos de referencia en las transcripciones hayan sido codificados y todas las frases de referencia hayan sido examinadas para garantizar que no queden conceptos indefinidos.

Generación de informes

Cuando se ha concluido cualquier proyecto investigativo, es indispensable otorgar a la población de estudio un informe con respecto a los resultados que surgieron de la investigación; esto permitiría a las personas participantes conocer sus fortalezas y debilidades en el contenido o en el constructo, e implementar programas de mejora. Igualmente, se demostraría la transparencia con respecto a la investigación realizada y, por ende, se ayudaría a motivar a las personas a participar en estudios futuros. Cabe destacar que, lo importante en la generación de los informes es el mantenimiento de la ética: ser honestos y respetuosos con respecto a los datos presentados, analizados y otorgados por los demás.

Evidencias de habilidades de razonamiento cuantitativo mediante la aplicación de los test *thinking aloud*

Definir el propósito de efectuar los test thinking aloud

Este estudio forma parte de una investigación principal que inició en el año 2020. El principal objetivo del estudio fue analizar las evidencias de validez, que permiten realizar inferencias sobre las habilidades de razonamiento cuantitativo demostradas por un grupo de examinados en la Prueba de Habilidades Cuantitativas y las requeridas en los cursos de Química General I e Introducción a la Química en la Universidad de Costa Rica.

Para cumplir con el objetivo principal, el estudio se dividió en siete etapas y en una de ellas se realizaron los test *thinking aloud* con la finalidad de obtener evidencias de validez de contenido y del constructo de la Prueba de Habilidades Cuantitativas (PHC), a partir de los procesos de respuesta realizado por un grupo de examinados y, de esta manera, identificar las habilidades de razonamiento cuantitativo demostrado por las personas a la hora de resolver los ítems.

Cabe mencionar que, la PHC es un requisito de admisión en carreras que lo solicitan y está formada por 40 ítems de selección única con cuatro opciones de respuesta. Cada año, desde el 2015, las personas se inscriben para luego postularse a las carreras que la tienen como requisito de admisión. Además, la dificultad TRI promedio es igual a 0, y se busca maximizar la precisión del nivel de habilidad (θ=0) (Rojas-Torres & Ordóñez-Gutiérrez, 2019).

Elaboración de los procesos de cada ítem

Para llevar a cabo los test *thinking aloud*, primeramente, se elaboró un marco teórico de referencia en el que se define el constructo razonamiento cuantitativo que se mide mediante la PHC. Además, dicho constructo se explica por las siguientes dimensiones: cuantificar, relacionar, clasificar, ejemplificar, validar y generalizar.

Luego, se solicitó a un grupo de expertos que resolvieran los ítems de la PHC que se aplicó en el año 2020, para los ingresos a las carreras en el año 2021. Cada experto los resolvió y los clasificó según las dimensiones propuestas del constructo. Además, se solicitó analizar si los ítems podían medir distintas dimensiones al mismo tiempo, según el proceso de solución realizado, o bien, el camino escogido para resolver el reactivo. Las soluciones fueron codificadas de acuerdo con las dimensiones del constructo.

Elaboración de los procesos de cada ítem

Para llevar a cabo los test *thinking aloud*, primeramente, se elaboró un marco teórico de

referencia en el que se define el constructo razonamiento cuantitativo que se mide mediante la PHC. Además, dicho constructo se explica por las siguientes dimensiones: cuantificar, relacionar, clasificar, ejemplificar, validar y generalizar.

Luego, se solicitó a un grupo de expertos que resolvieran los ítems de la PHC que se aplicó en el año 2020, para los ingresos a las carreras en el año 2021. Cada experto los resolvió y los clasificó según las dimensiones propuestas del constructo. Además, se solicitó analizar si los ítems podían medir distintas dimensiones al mismo tiempo, según el proceso de solución realizado, o bien, el camino escogido para resolver el reactivo. Las soluciones fueron codificadas de acuerdo con las dimensiones del constructo.

Elección del tamaño de la muestra

Para elegir a los participantes se consideró lo que teóricamente está establecido. Esto es, se escogieron las personas que realizaron la PHC en el año 2020, que ingresaron a alguna de las carreras en el año 2021 y que tuvieran en su malla curricular el curso Introducción a la Química o Química General I. Las personas seleccionadas estaban empadronadas para el año 2021 en alguna de las siguientes carreras: Farmacia, Química, Ingeniería Mecánica, Ingeniería Eléctrica, Física, Meteorología y Geología. Además, las personas seleccionadas obtuvieron una calificación mayor e igual a 90 en la PHC que se aplicó en el año 2020 para los ingresos a las carreras porque se quería sustentar la teoría y no establecer niveles de razonamiento.

La cantidad de participantes fue de 13 personas, 7 mujeres y 6 varones, la cual se consideró suficiente, ya que la finalidad fue comprobar la teoría y no su elaboración. Además, a partir del quinto participante, se logró determinar los patrones de respuesta en función de la teoría, tal y como mencionan Padilla y Benítez (2014).

Entrenamiento

Antes de que la persona participante realizara el proceso de los test *thinking aloud*, se leyó el consentimiento informado y después se brindó una explicación para llevar a cabo el entrenamiento. Para esto, la persona investigadora efectuó el protocolo de resolución para que la persona participante observara cómo debía proceder. Cada participante tuvo que resolver dos ítems que estaban en una ficha llamada *Ficha de entrenamiento para realizar la técnica thinking aloud*, la cual contenía dos ítems muy semejantes a los que se proporcionan en la PHC. Luego de resolver los ítems de entrenamiento, la investigadora determinaba si la persona participante lograba captar qué se requería en el ejercicio o no; si la persona lograba tener la fluidez en el ejercicio, entonces, iniciaba el test *thinking aloud* con los reactivos de la prueba meta, y si no, debía volver a realizar dicho ejercicio.

Es importante comentar que la mayoría de las personas participantes lograron captar, de manera inmediata, qué era lo que debían realizar en el primer intento; solamente una de las participantes tuvo que realizar el ejercicio dos veces, en gran medida debido a la ansiedad y nerviosismo que presentó. Ahora bien, cuando se evidenciaba exceso de nerviosismo, la investigadora procedió a entablar una conversación, sin que la persona participante se sintiera amenazada o estresada; después, la misma persona optaba por iniciar con los test *thinking aloud*.

Recolección de los datos

Con respecto a la recolección de los datos, se siguió lo establecido por Padilla y Leighton (2017), Fonteyn et al. (1993) y Green (1998), quienes indican que los test *thinking aloud* se deben realizar en un lugar tranquilo, que deben

efectuarse de manera individual y que se deben grabar en audio o en video. Para dicho proceso de recolección de los datos, se prepararon instrucciones haciendo explícito lo que se debía realizar. Además, los test *thinking aloud* se realizaron en las habitaciones propias de cada uno de los participantes porque a nivel país, se encontraban en confinamiento debido a la pandemia por COVID-19. También, se les explicó a las personas qué se esperaba de cada uno, el procedimiento y la duración de los test. Se realizó un entrenamiento antes de iniciar con los ítems del estudio y, cuando se generaban períodos de silencio, la investigadora les decía: "Recuerda que debes verbalizar mientras estás resolviendo los ítems".

Las verbalizaciones se realizaron, primeramente, de manera concurrente, y luego, de manera retrospectiva. En el caso del modo concurrente, las personas participantes debían verbalizar lo que estaban pensando en el momento de resolver el ítem, mientras que, para el modo retrospectivo, la persona debía explicar a la investigadora qué fue lo que empleó para resolver el ítem (Padilla & Leighton, 2017).

Transcripción, codificación y análisis de los thinking aloud

Luego de las verbalizaciones de los *test thinking aloud*, se procedió a la transcripción de las verbalizaciones de las 13 personas participantes y de los 40 ítems que contenía el formulario de la PHC. Las transcripciones se realizaron en *Word* y luego, se limpiaron quitando las verbalizaciones del enunciado de los ítems y solamente se dejó el proceso de respuesta del grupo de participantes.

Los análisis se realizaron mediante el software *Atlas ti.9* y se consideraron las postulaciones teóricas con respecto al constructo de razonamiento cuantitativo y sus dimensiones. Luego, se realizó un análisis de frecuencia de palabras. Posteriormente, se tomaron frases y un grupo de personas expertas en el constructo realizó un juzgamiento sobre los procesos de solución producidos por las personas participantes codificando y clasificando los procesos en cada una de las dimensiones del constructo.

Como cada persona experta debía juzgar el proceso de solución y clasificarlo, dicho juzgamiento se analizó mediante el índice de *Kappa de Fleiss*. De esta manera, se evaluó la estabilidad de acuerdos entre los jueces al asignar calificaciones categóricas a un ítem determinado. Cabe mencionar que, al grupo de jueces se le entregó un instrumento donde debía indicar: (*-1*) si el proceso efectuado por el conjunto de examinados no correspondía a lo solicitado en el ítem, según categoría; (*0*) si el proceso efectuado por el conjunto de examinados correspondía medianamente a lo solicitado en el ítem, según categoría; y (*1*) si el proceso efectuado por el conjunto de examinados correspondía fuertemente a lo solicitado en el ítem, según categoría.

Generación de informes

La generación de informes implica proporcionar los resultados del estudio. De esta manera, como primer resultado de la aplicación de los test *thinking aloud*, tenemos el análisis de la frecuencia de palabras que relacionan las verbalizaciones con lo establecido teóricamente. Esta frecuencia se establece en la Figura 1.

Se puede notar que la palabra más frecuente entre el grupo de personas participantes fue "calcular". Sin embargo, se debe aclarar que las personas, cuando están ante la resolución de una prueba que tiene contenido matemático, a todo le establecen el sello de "calcular". También, se pueden observar las palabras "relacionar", "patrón",

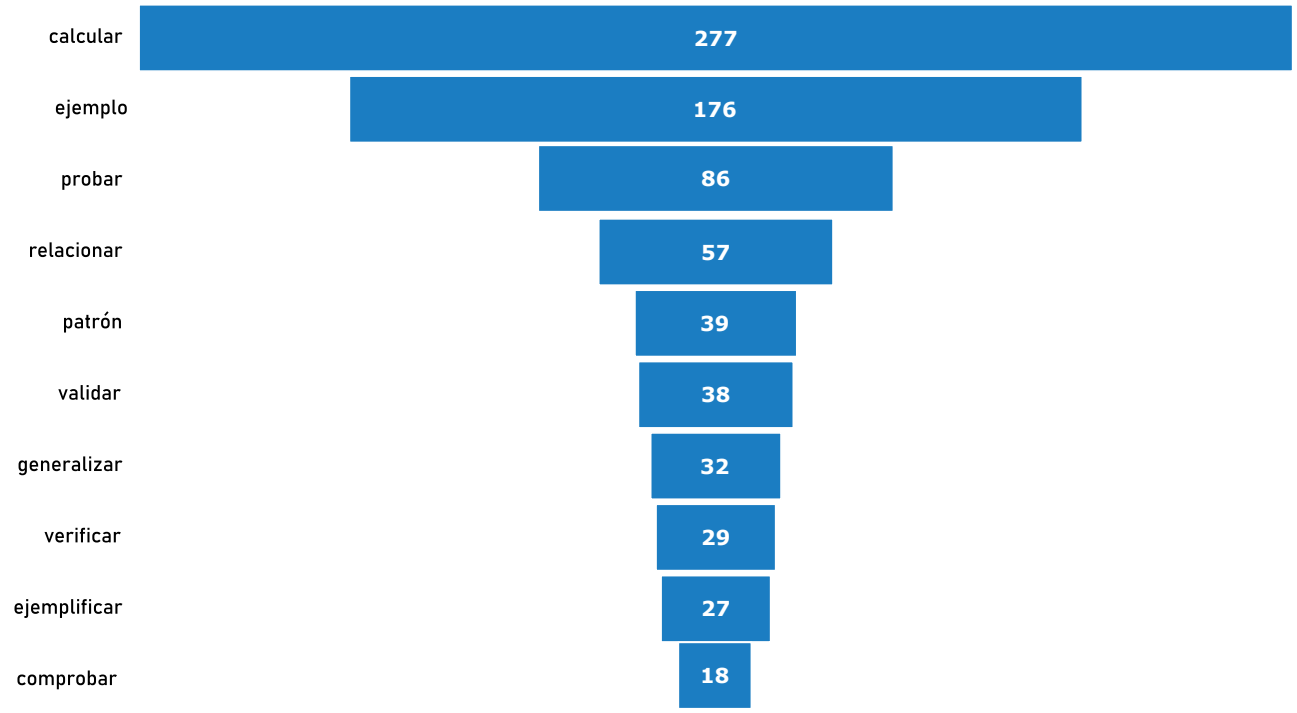


Figura 1
Análisis exploratorio de la frecuencia de palabras empleadas por el grupo de personas participantes en la resolución de los ítems de la PHC, 2020.

"ejemplo", "validar", "generalizar", "comprobar" y "verificar". Al revisar la literatura correspondiente a las dimensiones del razonamiento cuantitativo, se puede notar que las personas participantes emplean estas palabras en la resolución de los ítems y que, además, forman parte de las habilidades de razonamiento cuantitativo que se requieren en los cursos de Química indicados.

Como segundo resultado, se obtuvo la congruencia entre jueces sobre el análisis de los procesos de respuesta y su clasificación en las dimensiones del constructo. El índice de congruencia calculado fue *Kappa de Fleiss*, según la categoría del constructo y el formulario analizado, y se presenta en la Tabla 1.

Se puede notar que los valores más bajos corresponden a la categoría "relacionar". Sin embargo, Ruiz y Luciano (2012) explican que en la Teoría del Marco Relacional existen diferentes tipos de relaciones de acuerdo con los procesos que se deban emplear en la resolución de un problema, y que estas son: coordinación (es, es igual a, es semejante, es similar), oposición (es opuesto de, es lo contrario a, es diferente de), espaciales (arriba, abajo, lejos, cerca), deícticos (yo, tu, aquí, allí, antes, después), y que la visualización dependerá del contexto de formación de cada individuo. Ahora bien, el índice de congruencia sobre los procesos de solución para el formulario en general fue de 0.77, 0.75, 0.75 y 0.78 respectivamente y se consideró un valor muy bueno.

Por otro lado, se logró establecer cuáles habilidades de razonamiento cuantitativo lograban demostrar el grupo de examinados a la hora de resolver los ítems de la Prueba de Habilidades Cuantitativas y se tomó en cuenta la teorización del constructo y sus dimensiones. Estas se establecen a continuación en la Tabla 2.

Finalmente, se identificó en los patrones de respuestas aquellas habilidades que el grupo de

Tabla 1
Valor *Kappa de Fleiss* para la congruencia entre jueces según categorización de los ítems de la PHC 2020 por cada cuadernillo (o fórmula).

Categorías	*Kappa de Fleiss*			
	F1	**F2**	**F3**	**F4**
Calcular	0.73	0.72	0.72	0.73
Relacionar	**0.68**	**0.63**	**0.63**	**0.68**
Clasificar	1.00	1.00	1.00	1.00
Ejemplificar	0.85	0.85	0.85	0.85
Validar	0.79	0.76	0.76	0.78
Generalizar	1.00	1.00	1.00	1.00

Nota. **p*-value = 0, z =14.8.

personas participantes logró demostrar en resolución de los ítems de manera concurrente y retrospectiva. Este informe se les devolvió y se brindaron sugerencias de mejora.

Discusión

Este artículo muestra las etapas para realizar los test *thinking aloud* mediante una aplicación teórico-práctica en la determinación de habilidades de razonamiento cuantitativo que conlleva revisar rigurosa y sistemáticamente elementos indispensables del constructo a considerar y así, no generar vacíos a la hora de analizar los datos que se obtienen.

Ahora bien, es indispensable efectuar los test *thinking aloud* de manera planificada y con un objetivo claro. Además, es necesario tomar en cuenta todos los aspectos importantes para el estudio; por ejemplo, el tiempo destinado para la ejecución de los reportes verbales en los que tanto el investigador como los participantes tengan el tiempo suficiente para abarcar todos los puntos que se presentan en los ítems y de manera tal que no se vuelva una carrera por terminar la tarea otorgada.

Es importante indicar que, realizar los test *thinking aloud* al estudiantado permitirá que el cuerpo docente innove, estructure y realice mejoras en su labor docente. En consecuencia, conocer cómo piensan los estudiantes su materia permitirá saber la manera en que los conceptos son aprendidos. Igualmente, se genera conciencia pedagógica, y se desarrollan habilidades, en particular, habilidades de razonamiento cuantitativo que resultan imprescindibles para el funcionamiento adaptativo en la sociedad actual, tal y como lo plantea Stelzer et al. (2020).

Es importante mencionar que, en este escrito, se pone a disposición una guía general de las etapas que se deberían seguir, como mínimo, para efectuar los test *thinking aloud* de manera tal que permitan a los investigadores obtener evidencias de validez de contenido con respecto al constructo de interés. Por otro lado, se aporta una bibliografía a la que se puede acudir para efectuar una revisión en profundidad. La aplicación de los test *thinking aloud* realmente no es reciente; sin embargo, se ha desarrollado poco para obtener evidencias de validez de contenido. Además, las investigaciones en las que se mencionan no establecen una estructura evidente de los pasos por seguir para una obtención exitosa de los datos y de los análisis que se deberían realizar al respecto.

Por otra parte, las etapas plasmadas en este artículo no son estáticas, ya que pueden variar se-

Tabla 2
Habilidades de razonamiento cuantitativo demostrado por un grupo de personas en la resolución de los ítems de la PHC.

Dimensión	Subcategoría	Verbalización
Cuantificar	Obtención de medidas	"Lo que hice fue *sacar las medidas* y con esas medidas grandes fui partiendo para encontrar las medidas más chiquititas y así, saber las medidas de los lados" (Exa1). "Empecé a sacar medidas y al final, hice la resta" (Exa4).
	Verificación de las respuestas	"Primero, hice un mini cálculo para ver, como, para saber si es negativo o positivo; luego, verifico que eso se dé..." (Exa5).
	Simplificación de expresiones	"Voy a simplificar un poco esta expresión" (Exa12).
	Efectuar operaciones	"Está multiplicando esto, entonces, como veo que este producto de la suma " (Exa1).
Relacionar	Coordinación	"Agrupo términos semejantes..." (Exa4). "Me daban *relaciones* entre diferentes variables...como no había una *relación* directa entre..." (Exa2).
	Oposición	"Es lo *opuesto* a lo que me piden" (Exa7). "*R* sería *mayor a p*..." (Exa1).
	Espaciales	"Entonces, comprobamos la de arriba" (Exa11). "Multiplicando arriba y abajo"(Exa12).
	Equivalencias	"Bueno, por la equivalencia que me están dando" (Exa3). "Yo saqué la *equivalencia* de lo que es el" (Exa8).
	Conexiones o transformación de expresiones	"Estoy *comparando* la opción D con esto, a ver si puedo ver alguna *conexión* de este tipo…" (Exa2). "Es la misma regla de tres, pero…" (Exa3).
Clasificar	Identificación de características numéricas	"Tiene que haber un número par en la operación y eso va a hacer que el producto sea par" (Exa2).
	Clasificación de figuras	"Lo que se me ocurre es dibujar esta línea y verlo (sic) como 4 triángulos (Exa4).
	Ubicación de elementos dentro de un conjunto	"W pertenece con total certeza al intervalo" (Exa3). "Los términos que tienen Y se agrupan en" (Exa1).
	Ordenar y caracterizar objetos	"No me dicen nada de cómo están ordenados" (Exa13). "Cuando ordenamos los datos como el…" (Exa12).
Ejemplificar	Identificación de patrones	"Usé ejemplos y busqué excepciones" (Exa10). "Lo que hice fue coger un ejemplo, bueno, varios ejemplos para ver si lograba ver algún patrón al (sic) algo parecido…" (Exa9).
	Verificación de respuestas	"Lo confirmó con varios números para estar segura" (Exa2). "Fui probando cada uno de ellos [las opciones de respuesta] hasta llegar a…" (Exa7).
	Dar solución al problema planteado	"Se me ocurre que la manera más fácil es que se haga un ejemplo" (Exa2). "Utilicé ejemplos hipotéticos y analicé que todo dependía" (Exa10).
	Encontrar de manera directa la solución	"Lo que hice fue hacer casos para ver cuál era la solución" (Exa6). "Entonces si S es 100 por ejemplo en este caso" (Exa3).

Dimensión	Subcategoría	Verbalización
	Determinar el valor de verdad de un enunciado	"Voy a *poner un ejemplo* 1, 2, 3, 4, 5 porque cumplen las condiciones..." (Exa10). "Si asumiera que es un cero... pero esto no me indica que alguna de las respuestas sea cierta" (Exa9).
	Plantear relaciones	"Utilicé ejemplos específicos para ver la relación" (Exa6). "Por ejemplo, digamos que la x sea mayor que la y" (Exa1).
	Recordar conceptos	"Voy a darme valores para recordar..." (Exa4). "Estoy pensando en un ejemplo primero para ver..." (Exa5).
Validar	Obtención del valor de verdad mediante ejemplos específicos	"Eso no es cierto porque ya sabemos que es" (Exa5, Exa7).
	Conjeturar mediante juicios de valor	"*Bajo ese escenario*, las otras tres opciones se descartan y la C no" (Exa13). "Entonces, voy a volver a comprobar la opción A y la opción D" (Exa4).
	Confirmación de proposiciones	"Voy a confirmar con las otras opciones" (Exa1). "Se me ocurre cómo confirmar que eso puede ser" (Exa9).
Generalizar	Identificación de patrones	"Debería ver el patrón que sigue" (Exa2). "Aquí, lo que estoy haciendo es tratando (sic) de identificar un patrón"(Exa3).
	Encontrar una solución mediante un patrón establecido	"Cómo hago para generalizar el 4n" (Exa8). "Estoy intentando de (sic) encontrar la manera lógica de generalizar el n" (Exa13).
	Identificación de propiedades	"Veo un tipo de patrón en los resultados" (Exa13). "Me doy cuenta (sic) que todo producto de consecutivos debería ser la generalización de que todo..." (Exa2).

gún la finalidad de la investigación y el objetivo que se persigue con el instrumento de medición. Por ello, se debe tener en cuenta *qué* y *para qué* efectuar los test *thinking aloud*, además de saber cuál es el constructo para considerar. Igualmente, la manera de efectuar los análisis puede variar, ya que dependen de los avances en las tecnologías y de las nuevas propuestas que surjan para recolectar los datos de manera rigurosa, de manera tal que se obtengan mejores resultados al respecto.

Referencias

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

Artavia-Medrano, Á. (2015). Interpretación y análisis de pruebas psicológicas y educativas con el método Rule-Space. Revista *Actualidades en Psicología: Medición y Psicometría, 29*(119), 63-77. https://doi.org/10.15517/ap.v29i119.18724

Brizuela, A., Jiménez-Alfaro, K., Pérez-Rojas, N., Rojas-Rojas, G. (2016). Autorreportes verbales en voz alta para la identificación de procesos de razonamiento en pruebas estandarizadas. *Revista Costarricense de Psicología, 35*(1), 17-30. http://dx.doi.org/10.22544/rcps.v35i01.02

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

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). Wiley Blackwell.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming source of Differential Item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24-35. https://doi.org/10.1111/j.1745-3992.2010.00173.x

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* The MIT Press.

Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. En C. Faerch & G. Kasper (Eds.), *Multilingual matter: Introspection in second language research* (pp. 24-53). Multilingual Matters.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Fonteyn, M. E, Kuipers, B., & Grobe, S. J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research, 3*(4) 430-441. http://dx.doi.org/10.1177/104973239300300403

Green, A. (1998). *Verbal Protocol Analysis in Language Testing Research: A Handbook.* Cambridge University Press.

Joseph, G., & Patel, V. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making, 10*(1), 31-46. https://doi.org/10.1177/0272989X9001000107

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2017). Developing and validating cognitive models in assessment. En A. A. Rupp & J. P. Leighton (Eds.). *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications.* (pp. 75-101). Wiley Blackwell. https:/doi/10.1002/9781118956588.ch4

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https:/doi/10.2307/2529310

Leighton, J. P. (2004). Avoiding misconception, misuse and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(4), 6-15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education, 26*(2), 136-157. https://doi.org/10.1080/08957347.2013.765435

Leighton, J. P., Cui, Y., Ken-Cor, M. (2009). Testing expert-based and student-based cognitive models: An application of the attribute Hierarchy method and Hierarchy Consistency Index. *Applied Measurement in Education, 22*(3), 229-254. https://psycnet.apa.org/doi/10.1080/08957340902984018

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

Padilla, J. L, & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. En B. D. Zumbo, & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation Research* (pp. 211-228). Springer. https://doi.org/10.1007/978-3-319-56129-5_12

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136-144. https://doi.org/10.7334/psicothema2013.259

Prieto-Adánez, G. (2011). Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement. *Psicothema*, *23*(2), 233-238. http://www.psicothema.com/pdf/3876.pdf

Rojas-Torres, L., & Ordóñez-Gutiérrez, G. (2019). Proceso de construcción de pruebas educativas: El caso de la Prueba de Habilidades Cuantitativas. *Revista Evaluar, 19*(2), 15-29. https://doi.org/10.35670/1667-4545.v19.n2

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Tijdschrift voor Onderwijsresearch, 2*(2), 49-60. https://files.eric.ed.gov/fulltext/ED121845.pdf

Ruiz, F.J. & Luciano, C. (2012). Relacionar relacio-

nes como modelo analítico-funcional de la analogía y la metáfora. *Revista Latina de Análisis de Comportamiento,* 20. https://www.redalyc.org/articulo.oa?id=274525194014

Russo, J. E., Johnson, E. J. and Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*, 759 – 769. https://link.springer.com/article/10.3758/BF03202637#Bib1

Sapsirin, S. (2016). The application of verbal protocol analysis in second/foreign language testing research. Revista de revisión del idioma - *Universidad de Chulalongkorn,* 31. https://www.culi.chula.ac.th/en/pasaa-paritat/view/9

Stelzer, F., Vernucci, S., Aydmune, Y. S., del Valle, M. V., & Andrés, M. L. (2020). Diseño y validación de una escala de actitudes hacia las matemáticas. *Revista Evaluar, 20*(2), 51-68. https://doi.org/10.35670/1667-4545.v20.n2.30109

Van Den-Haak, M., De Jong, M., & Jan-Schellens, P. (2003). Retrospective vs. Concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339-351. http://dx.doi.org/10.1080/0044929031000

Virzi, R. A. (1992). Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? *Human Factors*, *34*(4), 457-468. https://doi.org/10.1177/001872089203400407