Actualidades Investigativas en Educación
ISSN: 1409-4703
ISSN: 1409-4703
Instituto de Investigación en Educación, Universidad de Costa Rica

Brizuela Rodríguez, Armel; Pérez Rojas, Nelson; Rojas Rojas, Guaner
Respuestas guiadas por el experto: Validación de las inferencias basadas en los procesos de respuesta
Actualidades Investigativas en Educación, vol. 18, núm. 3, 2018, , pp. 1-21
Instituto de Investigación en Educación, Universidad de Costa Rica

DOI: 10.15517/aie.v18i3.33456

Universidad de Costa Rica
Facultad de Educación
Instituto de Investigación en Educación

REVISTA ACTUALIDADES INVESTIGATIVAS EN EDUCACIÓN

http://revista.inie.ucr.ac.cr/ ISSN 1409-4703

Respuestas guiadas por el experto: Validación de las inferencias basadas en los procesos de respuesta

Expert-guided Responses: Validation of inferences based on response processes sciences"

Volumen 18, Número 3
Setiembre-Diciembre
pp. 1-21

Este número se publica el 1 de setiembre de 2018
DOI: https://doi.org/10.15517/aie.v18i3.33456

Armel Brizuela Rodríguez
Nelson Pérez Rojas
Guaner Rojas Rojas

Revista indizada en **REDALYC**, **SCIELO**

Revista distribuida en las bases de datos:

LATINDEX, **DOAJ**, **REDIB**, **IRESIE**, **CLASE**, **DIALNET**, **SHERPA/ROMEO**, **QUALIS-CAPES**, **MIAR**

Revista registrada en los directorios:

ULRICH'S, **REDIE**, **RINACE**, **OEI**, **MAESTROTECA**, **PREAL**, **CLACSO**

Respuestas guiadas por el experto: Validación de las inferencias basadas en los procesos de respuesta

Expert-guided Responses: Validation of inferences based on response processes

Armel Brizuela Rodríguez[1]
Nelson Pérez Rojas[2]
Guaner Rojas Rojas[3]

Resumen: *En este artículo se presenta un estudio cuantitativo cuyo objetivo fue poner a prueba un nuevo método para recabar evidencias sobre los procesos de respuesta utilizados por los examinados en pruebas educativas. Como antecedente fundamental, el estudio parte de la necesidad de contar con métodos sistemáticos para mejorar la calidad de los instrumentos que se utilizan en la evaluación educativa. Se comparó el método tradicional, de los reportes verbales, con el método de la Respuesta Guiada por el Experto con respecto a su idoneidad para recabar evidencias sobre la interpretación de las puntuaciones en una prueba educativa. Para esto, se seleccionó una muestra a conveniencia de 17 estudiantes de primer ingreso de la Universidad de Costa Rica y se le aplicó una entrevista semiestructurada a cada uno, en las cuales debían resolver ítems de razonamiento en voz alta empleando el método tradicional de reportes verbales o el método de la Respuesta Guiada por el Experto. Las entrevistas fueron codificadas por un grupo de expertos y se calculó un coeficente de acuerdo entre ellos. Con el método tradicional se obtuvo un coeficiente kappa de Fleiss de 0.22, mientras que, con el método de la Respuesta Guiada por el Experto, este fue de 0.40. Se concluye con las ventajas de utilizar este método para desarrollar instrumentos de evaluación educativa que representen adecuadamente las habilidades, destrezas, competencias y conocimientos de los estudiantes.*

Palabras clave: *evaluación educativa, estrategias de resolución, reportes verbales, acuerdo entre jueces.*

Abstract: *This article presents a quantitative study whose objective was to test a new method to collect evidence about the response processes used by examinees in educational tests. As a fundamental precedent, the study starts from the need to have systematic methods to improve the quality of the instruments used in educational assessment. A traditional method of verbal reports was compared to the Expert-guided Responses method regarding its suitability to gather evidence for scores interpretation in an educational test. For this purpose, a convenience sample of 17 first-year students from the University of Costa Rica was selected and a semistructured interview was applied to each, in which they had to solve reasoning items aloud using the traditional method of verbal reports, or the Expert-guided Responses method. The interviews were coded by a group of raters, and a coefficient of agreement between them was calculated. With the traditional method, a Fleiss kappa coefficient of 0.22 was obtained, whereas with the Expert-guided Responses method it was 0.40. The article concludes with the advantages of using the new method to develop educational assessment tools that adequately represent the students' abilities, skills, proficiencies, and knowledge.*

Key words: *educational assessment, solving strategies, verbal reports, inter-rater agreement.*

[1] *Es investigador y desarrollador de pruebas estandarizadas en el Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica. Máster en Investigación Psicológica de la misma Universidad. Dirección electrónica: armel.brizuelarodriguez@ucr.ac.cr*

[2] *Es investigador y desarrollador de pruebas estandarizadas en el Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica. Bachiller en Filología española de la misma Universidad. Dirección electrónica: nelson910@gmail.com*

[3] *Coordinador académico del Programa de la Prueba de Aptitud Académica del Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica. Doctor en Metodología de las Ciencias del Comportamiento y de la Salud de la Universidad Autónoma de Madrid, España. Dirección electrónica: guanerdavid@yahoo.com*

Artículo recibido*: 29 de noviembre, 2017*
Enviado a corrección*: 9 de marzo, 2018*
Aprobado*: 21 de mayo, 2018*

1. Introducción

La evaluación es un componente esencial en los procesos de enseñanza y aprendizaje. En este sentido, Segura (2009, p. 2) plantea que el propósito de la evaluación educativa es "la mejora constante del hecho educativo en un contexto humano social". Para ello, es necesario emplear diversos métodos e instrumentos que permitan valorar el desempeño mostrado por el estudiantado en cuanto a las habilidades, destrezas, competencias y conocimientos de interés para el evaluador.

Los instrumentos empleados en la evaluación educativa deben ser de alta calidad para cumplir con su propósito. Al respecto, Zapata y Canet (2008) argumentan que la construcción de este tipo de instrumentos implica una preocupación constante por aspectos como la validez y fiabilidad. Estos autores plantean un modelo de 11 pasos para la elaboración de un instrumento de evaluación, a saber: (1) Especificación del constructo, (2) definición de las dimensiones, (3) selección de los ítems y de la técnica de escalamiento, (4) validez de contenido, (5) diseño de la población y de la muestra, (6) prueba piloto, (7) ajuste a la escala, (8) aplicación del cuestionario, (9) diseño del diagrama de senderos, (10) fiabilidad y validez de constructo, y (11) ajuste final de la escala. No obstante, dentro de este modelo no se toma en cuenta la indagación sobre los procesos de respuesta de los estudiantes a la hora de contestar las preguntas de un instrumento de evaluación.

Uno de los métodos que existen para evaluar la calidad de un instrumento de evaluación educativa es la exploración de las estrategias y conocimientos que utilizan los estudiantes para contestarlo (Garrison y Andrews-Larson, 2016; Karabenick et al., 2007; Leighton y Gokiert, 2008; Ryan, Gannon-Slater y Culbertson, 2012; Smith, 2017). Este método permite identificar dificultades de comprensión de las preguntas, uso de estrategias no contempladas por el evaluador y otro tipo de problemas que ponen en duda la validez de las inferencias que se realizarán sobre los estudiantes con base en el resultado que obtengan en la prueba.

A pesar de que existen varias formas de implementar este método, estas tienen en común que (1) se basan en el reporte verbal en voz alta brindado por los estudiantes y (2) los expertos deben decidir si las habilidades, destrezas, competencias y conocimientos mostrados por estos corresponden a lo que realmente pretendían evaluar cuando elaboraron el instrumento. Asimismo, para que dicha exploración sea efectiva y cumpla con su propósito, es necesario que quienes la lleven a cabo (generalmente expertos en el área evaluada) evidencien un grado adecuado de consenso sobre las habilidades, destrezas,

competencias o conocimientos observados en los reportes verbales de los estudiantes. En consecuencia, el desarrollo de métodos que promuevan un nivel adecuado de consenso entre los jueces expertos es de suma importancia para mejorar por la evaluación educativa y con ello los procesos de enseñanza y aprendizaje.

Así pues, el objetivo de este artículo es presentar evidencias empíricas sobre la utilidad de un nuevo método para recabar evidencias sobre los procesos de respuestas para la validez de las interpretaciones de las puntuaciones de las pruebas educativas. De acuerdo con los estándares de la *American Educational Research Association* (AERA), la *American Psychological Association* (APA) y el *National Council on Measurement in Education* (NCME), publicados en 2014:

> some construct interpretations involve more or less explicit assumptions about the cognitive processes engaged in by test takers. Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or the response actually engaged in by test takers. (AERA, APA y NCME, 2014, pp. 14-15)

Para recabar este tipo de evidencias, existen diferentes métodos diseñados para analizar las estrategias de respuesta individuales en tareas cognitivas, tales como la entrevista cognitiva (Miller, Willson, Chepp y Padilla, 2014; Willis, 2015), los protocolos de pensamiento en voz alta (Fox, Ericsson y Best, 2011), el seguimiento ocular (Gorin, 2006) y el registro de los tiempos de respuesta (Sireci et al., 2008). De todas estas posibilidades para recabar evidencias de validez relacionadas con los procesos de respuesta, los reportes verbales en voz alta (i.e., las entrevistas cognitivas y los protocolos de pensamiento en voz alta) son los más utilizados debido a su relativo bajo costo económico y a su facilidad de uso (ver Figura 1). En este sentido, el seguimiento ocular y el registro de los tiempos de respuesta requieren de equipo especializado, cuyo manejo y costo económico los colocan fuera del alcance de muchos investigadores.

Así pues, para recabar evidencias de los procesos de respuesta en los ítems de elección única, generalmente se utilizan entrevistas individuales semiestructuradas en las que el investigador proporciona los ítems a una muestra pequeña de la población meta. Lo que se requiere de los participantes es (1) un reporte verbal concurrente (i.e., simultáneo), mientras intentan resolver los ítems o (2) un reporte verbal retrospectivo mediante el cual los participantes, una vez elegida la respuesta, describen qué estrategias o procesos siguieron

para escogerla. Este proceso puede ser más o menos guiado, ya que en algunos casos se permite que el participante reporte todo lo que considere pertinente, pero, en otros, el investigador guía el proceso a través de preguntas que apuntan a características específicas de los ítems. Posteriormente, los reportes verbales se transcriben y dos o más jueces los codifican de manera independiente para determinar si los procesos evidenciados durante la entrevista son consistentes con el constructo que se pretende medir con los respectivos ítems.

Figura 1
Instrucciones iniciales de una entrevista cognitiva

Instrucciones
Estamos realizando una serie de pruebas para ver cómo funciona este cuestionario. Para eso yo le voy a dar el cuestionario y le voy a pedir que lo llene como si estuviéramos realizando la encuesta. En esta etapa lo que nos interesa es saber cómo está funcionando el cuestionario. Por eso le voy a pedir que conforme lo va completando ***piense en voz alta***. Es decir, que diga en voz alta todo lo que se le viene a la mente conforme va completando las preguntas.
En cada pregunta, *yo le voy a realizar más preguntas sobre la redacción de estas,* las instrucciones y las opciones de respuesta. Es importante que tenga presente que nosotros queremos saber si el cuestionario funciona. No dude en decirme si algo le parece confuso y si algo se puede mejorar.
Vamos a durar aproximadamente ____ minutos en todo el proceso.
Antes de iniciar ¿tiene alguna pregunta?
Práctica: Para irse acostumbrando a pensar en voz alta, vamos a practicar con la primera sección del cuestionario. A partir de la segunda sección vamos a iniciar formalmente con la entrevista.

Fuente: Elaboración de la Dra. Vanessa Smith Castro y el Dr. Mauricio Molina, 2011.

Este método, para indagar sobre la forma en la que los examinados se enfrentan a los ítems, ha sido de gran utilidad para diversos propósitos, tales como explorar habilidades de resolución de problemas en biología y en ciencias políticas (Taylor y Dionne, 2000) o de comprensión lectora (Cromley y Azevedo, 2006), validar modelos cognitivos subyacentes en pruebas educativas (Leighton y Gierl, 2007; Wang y Gierl, 2011), examinar fuentes de funcionamiento diferencial de los ítems (Ercikan et al., 2010), recabar evidencias de validez sobre el funcionamiento de pruebas psicológicas (Castillo y Padilla, 2013), explorar las posibles causas del desajuste entre un patrón de respuestas y un determinado modelo

psicométrico (Cui y Roduta, 2013), entre otros. Como se puede apreciar, los reportes verbales son una técnica polifacética y de gran versatilidad que ha permitido generar información sumamente valiosa sobre el funcionamiento de las pruebas educativas.

Por ejemplo, Farr, Pritchard y Smitten (1990), a través de este método, pudieron identificar que, a pesar de la gran variabilidad de estrategias utilizadas para resolver una prueba educativa, los examinados se enfocaban especialmente en revisar primero los ítems y luego proceder a leer el texto necesario para responderlos. En este caso, los reportes verbales aportaron una información de gran relevancia para los constructores de pruebas, la cual es complementaria a la puntuación total obtenida por los examinados o a los patrones de opciones seleccionadas por estos.

En otro estudio, Powers y Leung (1995) pudieron establecer a partir de entrevistas realizadas a un grupo de examinados que era posible obtener puntuaciones elevadas en el apartado de comprensión lectora del *Scholastic Aptitude Test* sin leer los textos que, en principio, son necesarios para contestar las preguntas de esta prueba. El método de los reportes verbales les permitió observar que la estrategia más utilizada era revisar la consistencia lógica entre las preguntas y reconstruir el tema original de estas. Con base en esto, los investigadores concluyeron que las estrategias de resolución de esta prueba no son indicadores de comprensión lectora. En este sentido, los reportes verbales sirvieron como una estrategia para cuestionar la validez de las inferencias sobre el nivel de comprensión lectora de los examinados, lo cual responde directamente a los requerimientos de los estándares de la AERA, la APA y el NCME (2014).

Finalmente, es relevante mencionar el aporte de Rupp, Ferne y Choi (2006),estos investigadores analizaron la información registrada mediante 10 entrevistas cognitivas aplicadas a lectores no nativos del inglés, quienes debían contestar varios ítems de selección única basándose en tres textos extraídos del *Canadian Test of English for Scholars and Trainees*. Los autores identificaron un conjunto de estrategias para abordar los ítems, a saber: 1) revisar superficialmente el texto, después las preguntas y posteriormente buscar términos importantes en el texto principal para seleccionar la alternativa correcta; 2) revisar todo el texto, luego buscar palabras clave, después leer las alternativas, identificar términos clave en las opciones y responder los ítems; c) leer los ítems de primero, posteriormente, el pasaje en el que se basan los ítems, y escoger la opción correcta; y d) revisar primero las preguntas para después tratar de encontrar palabras importantes en el texto principal e ir simultáneamente seleccionando las alternativas correctas en función de

las palabras que iban encontrando. Al igual que en los estudios mencionados anteriormente, el interés por analizar los reportes verbales de los examinados permitió identificar procesos de respuesta relevantes, los cuales deben ser tomados en cuenta por los desarrolladores de las pruebas educativas, en aras de garantizar la validez de las inferencias realizadas.

Con base en esta breve reseña, es posible concluir que los reportes verbales en sus diferentes variedades representan una herramienta de gran utilidad para generar evidencias sobre los procesos de respuesta empleados en las pruebas educativas, por cuanto permiten identificar fuentes de variabilidad ajenas a las habilidades, destrezas, competencias o conocimientos de interés (Castillo y Padilla, 2013; Cui y Roduta, 2013; Ercikan et al., 2010; Taylor y Dionne, 2000). Lo anterior se cumple en el tanto en el que el método se utilice correctamente (Cromley y Azevedo, 2006; Leighton, 2004) y en un contexto relevante (Leighton y Gierl, 2007).

Pese a sus evidentes beneficios, los reportes verbales, en voz alta, han sido cuestionados debido a las posibles interferencias en el proceso de resolución que implica el acto de intentar reportarlo en voz alta (Padilla y Benítez, 2014). Asimismo, otra desventaja de los reportes verbales es que deben ser codificados por jueces una vez que han sido transcritos, lo cual implica un componente de subjetividad difícil de controlar. Generalmente, los jueces son entrenados en el uso de los códigos para evitar discordancias entre ellos; sin embargo, estas generalmente persisten y se visualizan fácilmente cuando se realizan los respectivos cálculos de concordancia entre jueces. Esta desventaja se agrava si se toma en cuenta que los participantes muestran diferencias individuales en cuanto a su capacidad para reportar verbalmente qué estrategias emplean para contestar los ítems, por lo cual es frecuente que los jueces expertos deban codificar reportes verbales sumamente escuetos que incrementan aún más las discordancias entre ellos.

Al respecto, cabe señalar que el procedimiento usual para resolver estos problemas de desacuerdos entre los expertos es mediante una o varias reuniones en las que se discuten las discordancias y se llega a un acuerdo. No obstante, en la gran mayoría de estudios no se especifican con claridad los procedimientos seguidos para afrontar las discordancias entre los jueces. Lo usual es un reporte de unas cuantas líneas en las que se indica que los desacuerdos fueron resueltos después de una discusión entre los jueces, por lo cual no es posible descartar la posibilidad de que durante dicha discusión se haya introducido en la codificación una gran cantidad de sesgos asociados a las dinámicas de grupo específicas de dichas reuniones.

Ante este panorama, el presente estudio plantea un nuevo método para validar las inferencias sobre los procesos de respuesta en pruebas educativas, lo cual permite minimizar el problema de la concordancia entre jueces. La ventaja de este método es que no requiere que los participantes reporten en voz alta sus procesos de respuesta, por lo que simplifica, en gran medida, la labor de codificación de los jueces y, por ende, minimiza los desacuerdos propios de este tipo de tareas cognitivas complejas. En el siguiente apartado, se presenta en qué consiste este método y cómo puede ser utilizado para recabar evidencias relacionadas con los procesos de respuesta en una prueba educativa.

2. El método de las Respuestas Guiadas por el Experto

Este método está diseñado para recabar evidencias sobre los procesos de respuesta en el caso de las pruebas de selección única, cuyos ítems muestran una estructura tripartita: (1) un enunciado principal que presenta información sobre alguna persona, cosa, animal o situación; (2) una pregunta que debe ser contestada con base en lo que plantee el enunciado principal; y (3) dos o más opciones de respuesta entre las que se debe seleccionar aquella que incluye la respuesta correcta a la pregunta. A partir de este modelo básico, se han generado diversas variaciones, por ejemplo, ítems con opciones parcialmente correctas o ítems en los que se plantean varias preguntas a partir de un mismo enunciado principal.

El uso del método de las Respuestas Guiadas por el Experto consiste en presentar una versión modificada del ítem a los participantes, en la cual no se incluyen las opciones de respuesta, sino solamente el enunciado principal y una pregunta que proporcione la información necesaria para llegar a una respuesta similar a la opción correcta del ítem en su versión original. Ahora bien, esta pregunta que incluye la información requerida para resolver el ítem será justamente una guía por parte del experto encargado de construir los ítems, quien previamente debe conocer (1) las estrategias o conocimientos requeridos para resolverlos y (2) la opción correcta del ítem en su versión no modificada. Así pues, si el examinado desarrolla (de manera oral o escrita) una respuesta similar a la opción correcta del ítem, se puede concluir que existe una evidencia inicial de que la información o guía aportada por el experto en la pregunta es relevante para resolver correctamente el ítem. Como último paso, al participante se le muestra la versión original del ítem y se le solicita que elija la opción correcta.

A modo de ejemplo, se presenta un ítem del libro *Resolvamos la PAA* (Brizuela et al., 2015), en el cual se presentan diversos ejemplos de ítems incluidos en la Prueba de Aptitud Académica para la admisión a la Universidad de Costa Rica y a la Universidad Nacional de Costa Rica. Posteriormente, se presenta este ítem modificado, el cual sería mostrado a los participantes de una investigación sobre los procesos de respuesta para resolver dicho ítem, y cuya pregunta se plantea de modo que incluya la información más relevante para resolverlo.

Figura 2
Versión original del ítem

Al ser humano debemos entenderlo como persona y como miembro de un grupo.

Según el texto anterior, el humano es un ser

A) individual y social.
B) biológico e integral.
C) unificado y complejo.
D) particular y universal.
E) anatómico y psicológico.

Fuente: Elaboración propia, con información del Programa de la Prueba de Aptitud Académica de la Universidad de Costa Rica, 2017.

En la Figura 1, se presenta un ítem con el que se pretende medir la capacidad del examinado para parafrasear correctamente el enunciado principal (Brizuela, Jiménez, Pérez y Rojas, 2016). Por lo tanto, una evidencia de validez sobre los procesos de respuesta en este ítem sería el hecho de que el examinado pudiera desarrollar una respuesta similar a la opción correcta (la A) si se le proporcionara una guía o "pista" que, en este ejemplo, consistiría en indicarle que parafrasee lo dicho sobre el ser humano. Como se puede apreciar en la Figura 2, la versión modificada carece de opciones de respuesta, ya que el supuesto es que elegir la respuesta correcta no depende del formato del ítem, sino que se basa en utilizar la estrategia adecuada (en este caso, parafrasear).

Figura 3
Versión modificada del ítem

Al ser humano debemos entenderlo como persona y como miembro de un grupo.

Mencione dos calificativos que expresen la misma idea del enunciado en relación con el ser humano.

Fuente: Elaboración propia, con información del Programa de la Prueba de Aptitud Académica de la Universidad de Costa Rica, 2017.

Una vez que al participante se le muestra la versión modificada del ítem, se le solicita que conteste la pregunta de manera oral o escrita. Posteriormente, se le proporcionan, por escrito, las opciones de la versión original para que selecciona la correcta. Si las respuestas de los participantes fueron orales, será necesario transcribirlas para los análisis que se explican a continuación.

Cuando se cuenta con el registro escrito de las respuestas de los participantes, estas deben ser codificadas de manera independiente por, al menos, dos jueces. Su tarea consiste en determinar si la respuesta desarrollada por el participante es similar a la opción correcta del ítem en su versión original. Para esta codificación, los jueces pueden utilizar una escala dicotómica (Sí/No) o politómica (Diferente/Cierto, parecido/Similar), de modo que finalizado el proceso de codificación sea posible calcular algún índice de concordancia entre jueces, como el coeficiente kappa de Cohen (Landis y Koch, 1977) o algún índice de correlación intraclase (Shrout y Fleiss, 1979).

Así pues, en el presente artículo se compara el método tradicional de los reportes verbales con el método de las Respuestas Guiadas por el Experto, a fin de presentar evidencias empíricas sobre su utilidad para recabar evidencias sobre los procesos de respuesta utilizados en pruebas educativas.

3. Método

Este es un estudio cuantitativo con un enfoque exploratorio y cuyo alcance es descriptivo, en el cual se aportan evidencias preliminares sobre la utilidad de un nuevo método para recopilar evidencias relacionadas con los procesos de respuesta empleados en pruebas educativas.

3.1 Participantes

Se seleccionó un grupo de 17 personas que ingresaron en el año 2016 a la Universidad de Costa Rica. Estas personas fueron elegidas debido a que obtuvieron, al menos, un 90% de preguntas correctas en la Prueba de Aptitud Académica y por provenir de una institución pública de educación secundaria. Este criterio de inclusión, basado en la proporción de respuestas, se realizó con la intención de seleccionar participantes que tuvieran una mayor facilidad para resolver los ítems y para reportar en voz alta las respuestas.

En lo concerniente al aspecto ético, la ejecución de la presente investigación fue aprobada por el Comité Ético Científico de la Universidad de Costa Rica en la sesión N°10 del 27 de abril de 2016. Dada la naturaleza observacional, no biomédica y de nulo riesgo, a los participantes solamente se les solicitó su consentimiento oral antes de iniciar las sesiones.

3.2 Instrumentos

Se utilizaron 14 ítems del área verbal de la Prueba de Aptitud Académica, los cuales forman parte del banco de ítems empleados regularmente para el ensamblaje de esta prueba, por lo cual se consideran material confidencial y no pueden ser mostrados en el presente artículo. Estos 14 ítems fueron seleccionados debido a que fueron clasificados previamente por los investigadores a cargo del estudio como indicadores de las estrategias de resolución definidas en la Tabla 1, las cuales fueron desarrolladas como parte del estudio publicado en Brizuela et al., (2016).

Los ítems empleados tienen un formato de selección única, con cuatro opciones de respuesta incorrectas y una correcta. Dichos ítems se refieren a situaciones cotidianas que son parte de las experiencias (educativas, personales, laborales, imaginarias, etc.) a las que comúnmente se enfrentan los aspirantes, y su objetivo es medir las habilidades de razonamiento en contextos verbales para propósitos de ingreso a la Universidad de Costa Rica y a la Universidad Nacional de Costa Rica (Brizuela y Montero, 2013; Brizuela et al., 2016; Jiménez y Morales, 2009-2010; Rojas, 2013; Rojas, 2014).

Tabla 1
Estrategias de resolución

Estrategia	*Definición*
Suponer	Identificar la única conclusión verdadera que puede inferirse asumiendo que las premisas dadas e implícitas en el texto principal son verdaderas.
Presuponer	Reconocer una relación pragmática entre las posibles denotaciones y connotaciones de ciertas palabras del texto principal y de las opciones de respuesta.
Reducir	Identificar la síntesis de la información expuesta en el texto principal en una proposición o descubrir un elemento en común entre los significados de varias palabras, frases u oraciones.
Oponer	Percatarse de que una o dos palabras del texto principal son antónimas (u opuestas en algún sentido) entre sí o respecto de la respuesta correcta.
Parafrasear	Percatarse de que una o dos palabras del texto principal son sinónimas (u equivalentes en algún sentido) entre sí o respecto de la respuesta correcta.

Fuente: Elaboración propia, con información de Brizuela et al. (2016).

3.3 Procedimiento

Los participantes fueron reclutados por teléfono y por correo electrónico para que asistieran a las oficinas del Programa de la Prueba de Aptitud Académica. A 11 participantes se les entregó una tarjeta con la versión modificada de cada ítem (ver Figura 2), con un enunciado principal y una pregunta en la que se guiaba al participante hacia una estrategia apropiada para contestar correctamente. Una vez que el participante leía y respondía la pregunta, se le entregaba otra tarjeta en la que se incluían las opciones de respuesta originales de cada ítem (ver Figura 1), con el objetivo de que seleccionara la alternativa correcta. A los restantes 6 participantes, solamente se les entregó las versiones originales de los ítems y se les solicitó que los resolvieran en voz alta. Las sesiones con cada participante fueron registradas mediante una grabadora de audio. Estas grabaciones fueron posteriormente transcritas por dos de los investigadores a cargo del presente estudio.

Una vez transcritas las respuestas de los participantes, cuatro jueces se encargaron de codificarlas de manera independiente, estas personas eran los desarrolladores de los ítems verbales de la Prueba de Aptitud Académica. Para esta labor de codificación, los jueces emplearon dos plantillas (Figura 4 y Figura 5), cuyas filas correspondían al número de ítem que contestó cada participante. En el caso de las 11 sesiones en donde se empleó el método de las Respuestas Guiadas por el Experto, los jueces solamente debían escribir si la respuesta brindada por los participantes al ítem modificado era similar a la respuesta correcta de la versión no modificada de dicho ítem, para lo cual los jueces disponían de las versiones originales de los ítems. Los códigos ("Sí" o "No") asignados por los jueces a cada respuesta fueron tabulados en una base de datos digital.

Por otra parte, en el caso de las 6 sesiones en las que se utilizó el procedimiento tradicional de los reportes verbales en voz alta, los jueces debían codificar qué tipo de categoría o estrategia de resolución se evidenciaba en el reporte verbal de los participantes. Algunos ejemplos de las categorías empleadas por los jueces pueden consultarse en Brizuela et al. (2016), estudio en el que también se explica detalladamente cómo se elaboraron las categorías.

Figura 4
Plantilla empleada por jueces con ítems modificados

Número de grabación	
Ítem	¿La respuesta sin opciones (respuesta abierta) es similar a la clave del ítem con opciones (selección única)?
(1)	
(2)	
(3)	

Fuente: Elaboración propia, con información del Programa de la Prueba de Aptitud Académica de la Universidad de Costa Rica, 2017.

Como se puede observar en la Figura 4, el proceso de codificación requerido implica un juicio de similitud entre dos o más proposiciones, mientras que en la Figura 5 se puede apreciar que la labor demandada al experto implica una mayor complejidad. En este sentido, la pregunta clave para los jueces, empleada en el método de las Respuestas Guiadas por el Experto, ayuda a reducir la subjetividad inherente a todo juicio emitido por un experto.

Figura 5
Plantilla empleada por jueces con versiones originales de ítems

Número de grabación	
Ítem	¿Cuál estrategia se evidencia en el reporte verbal del estudiante?
(1)	
(2)	
(3)	

Fuente: Elaboración propia, con información del Programa de la Prueba de Aptitud Académica de la Universidad de Costa Rica, 2017.

3.4 Análisis

Se realizaron análisis descriptivos para calcular el número de respuestas correctas para cada ítem, así como también se calculó la concordancia global entre los cuatro jueces mediante la implementación del coeficiente kappa de Fleiss del paquete estadístico STATA 14. Dado que el enfoque del estudio es exploratorio y con un alcance descriptivo, no se estimaron parámetros poblacionales ni se calcularon pruebas de significancia estadística.

4. Resultados

Dado que el objetivo del estudio fue poner a prueba un nuevo método para recabar evidencias de validez sobre los procesos de respuesta utilizados en pruebas educativas, el apartado de resultados se dividió en dos secciones. En la primera, se brindan los resultados observados con el método de las Respuestas Guiadas por el Experto, mientras que en la segunda se muestran los resultados correspondientes a la aplicación del método tradicional de los reportes verbales. De esta manera, es posible comparar fácilmente la idoneidad de ambos en términos de la concordancia entre jueces alcanzada.

4.1 Método de las Respuestas Guiadas por el Experto

Como se puede observar en la Tabla 2, para un total de 154 respuestas (11 participantes x 14 ítems = 154) que podían ser calificadas como correctas o incorrectas, solamente el 9,74% de estas fueron incorrectas y la proporción de respuestas correctas por ítem osciló entre 0,55 y 1,00. Este resultado se adecua a lo esperado de acuerdo con el método de las Respuestas Guiadas por el Experto, ya que en todos los casos la pregunta de los ítems incluía información estratégica y relevante para contestarlos correctamente. Sin embargo, debe recordarse que los participantes en este estudio fueron seleccionados porque mostraron un buen desempeño en la Prueba de Aptitud Académica que se aplicó en el año 2015 para el proceso de admisión 2015-2016. Por lo tanto, estos resultados deben interpretarse como evidencia preliminar de que la guía aportada por el experto fue relevante para acertar los ítems, pero no se puede dejar de lado que una hipótesis alternativa por descartar sería que acertaran los ítems debido a su alto nivel de habilidad de razonamiento y no a la guía aportada por el experto.

Por otra parte, tal y como se explicó anteriormente, cuatro jueces codificaron las respuestas orales transcritas de los participantes en términos de si estas eran similares o no a la alternativa correcta de la versión original del ítem. El cálculo del coeficiente kappa de Fleiss, para los cuatro jueces, arrojó el valor de 0,409, el cual se considera como moderado de acuerdo con Landis y Koch (1977). Este valor es sumamente satisfactorio, tomando en consideración que no hubo ningún tipo de conversaciones entre los jueces sobre el proceso de codificación. Contrario a lo que suele suceder cuando se emplea el método de los reportes verbales, el coeficiente reportado no se debe a que los jueces hayan llegado a un consenso en una reunión.

Tabla 2
Respuestas calificadas por participante y por ítem

Participante	It1	It2	It3	It4	It5	It6	It7	It8	It9	It10	It11	It12	It13	It14
Part 1	1	1	0	1	1	1	1	1	1	1	0	1	1	0
Part 2	1	1	1	0	0	1	1	1	1	1	1	1	1	1
Part 3	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Part 4	1	1	1	1	1	1	1	1	1	1	0	1	1	1
Part 5	1	1	1	0	1	1	1	1	1	1	1	1	1	1
Part 6	1	1	1	1	1	1	1	1	1	1	0	1	1	0
Part 7	1	1	0	1	1	1	1	1	1	1	1	0	1	1
Part 8	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Part 9	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Part 10	1	1	1	1	1	1	1	1	1	1	0	1	1	0
Part 11	1	1	1	1	1	1	0	1	1	1	0	1	1	1
Índice de dificultad	0,77	0,69	0,83	0,70	0,68	0,54	0,74	0,69	0,62	0,73	0,77	0,66	0,77	0,54
Proporción de de respuestas correctas	1,00	1,00	0,82	0,82	0,91	1,00	0,91	0,00	1,00	1,00	0,55	0,91	0,00	0,73

Fuente: Elaboración propia, con información del Programa de la Prueba de Aptitud Académica de la Universidad de Costa Rica, 2017.
Nota: El valor "1" corresponde a respuesta correcta y "0", a respuesta incorrecta. El índice de dificultad es el original del ítem, mientras que la proporción de respuestas correctas se obtuvo a partir de los 11 participantes del estudio.

4.2 Método tradicional de los reportes verbales

En la Tabla 3 se presenta la cantidad de veces que un determinado reporte verbal sobre un ítem fue clasificado por los jueces en las posibles estrategias para resolverlo. Se debe tomar en cuenta que, al tratarse de 6 participantes y 4 jueces, una concordancia perfecta se reflejaría si en una determinada fila de la Tabla 3 se observara el valor de 24 en una casilla y el valor de 0 en todas las demás. Aunado a ello, la columna encabezada por un "No responde" representa el número de ocasiones que los jueces no pudieron llevar a cabo una clasificación debido a que el reporte verbal del participante fue insuficiente para permitirles decidir a cuál estrategia asignarlo.

Así pues, la Tabla 3 muestra un conteo de las ocasiones en que los reportes verbales fueron clasificados en cada estrategia y para cada ítem. Las filas de la Tabla 3 corresponden a los ítems empleados en este estudio, de manera que se pueda apreciar la variabilidad de las clasificaciones de los reportes verbales realizadas por los expertos. Por ejemplo, para el ítem 1, los jueces consideraron los reportes verbales de los participantes como indicadores de la estrategia Suponer en 10 ocasiones, de la estrategia Oponer en 1 ocasión, de la

estrategia Parafrasear, en 8 ocasiones, y en 9 ocasiones no asignaron el reporte a ninguna estrategia de resolución.

En la Tabla 3 se puede apreciar no solo que hay una gran discordancia entre los jueces en cuanto a la clasificación de los reportes verbales, sino que además se observa una cantidad importante de valores en la columna que representa las ocasiones en las que los jueces no pudieron tomar una decisión sobre cuál estrategia utilizó el participante para resolver los ítems. Este patrón se refleja en que el cálculo del coeficiente kappa de Fleiss, para los cuatro jueces, arrojó el valor de 0,22, el cual, aunque se considera como aceptable de acuerdo con Landis y Koch (1977), es inferior al obtenido con la aplicación del método de las Respuestas Guiadas por el Experto.

Estos resultados son esperables, pues el método tradicional de los reportes verbales implica un juicio global de lo que decidan expresar en voz alta los participantes. En este sentido, para llevar a cabo la codificación de los reportes verbales, los expertos dependen de que los participantes posean una gran habilidad de verbalización, mientras que eso no ocurre en el caso del método de las Respuestas Guiadas por el Experto.

Tabla 3
Número de reportes verbales por ítem según estrategia de resolución

Ítem	Estrategia de resolución					
	Suponer	Presuponer	Reducir	Oponer	Parafrasear	No responde
1	10	0	0	1	8	5
2	11	7	0	0	0	6
3	6	3	0	0	4	11
4	4	19	0	0	0	1
5	6	8	2	1	2	5
6	12	3	2	0	0	7
7	8	13	0	0	1	2
8	9	8	1	0	4	2
9	5	3	9	0	6	1
10	6	2	8	1	4	3
11	2	10	6	0	0	6
12	3	1	14	4	1	1
13	9	0	9	0	2	4
14	14	3	3	0	2	2

Fuente: Elaboración propia, con información del Programa de la Prueba de Aptitud Académica de la Universidad de Costa Rica, 2017.

5. Conclusiones

Los instrumentos de evaluación educativa son herramientas de gran importancia que facilitan la monitorización de los aprendizajes alcanzados por los estudiantes. De ahí que, por ejemplo, no es de extrañar que, en el apartado 33 de *Educación 2030: Declaración de Incheon y Marco de Acción para la realización del Objetivo de Desarrollo Sostenible 4* (UNESCO, 2015, p. 37), se recalque la necesidad de "un entendimiento común y estrategias viables para evaluar el aprendizaje de maneras que garanticen que todos los niños y jóvenes, sin importar su situación, reciban una educación de calidad y pertinente, entre otras cosas sobre derechos humanos, arte y ciudadanía". Por este motivo, son ampliamente utilizados a nivel internacional para comparar el desempeño de los estudiantes en relación con estándares educativos de aprendizaje.

No obstante, a menudo las evaluaciones educativas son acusadas de mostrar sesgos culturales (Wagner, 2011) que amenazan la validez de las interpretaciones basadas en sus resultados. En este sentido, el contexto sociocultural en el que una persona nace y se desarrolla tiene un impacto crucial en los procesos de aprendizaje, el cual interactúa de diversas maneras con las capacidades cognitivas de los estudiantes (Basterra, Trumbull y Solano-Flores, 2011). Por lo tanto, cuando se desarrollan pruebas educativas es necesario garantizar que los distintos grupos culturales que conforman la población meta gocen de la misma oportunidad para evidenciar sus conocimientos y habilidades. Para ello se requiere que los desarrolladores de las pruebas educativas cuenten con métodos adecuados para identificar y analizar, de manera confiable y consistente, las habilidades, destrezas, competencias y conocimientos utilizados por los estudiantes cuando contestan las preguntas de un instrumento de evaluación.

En el presente estudio se presentaron evidencias empíricas sobre la utilidad del método de las Respuestas Guiadas por el Experto (RGE) para recabar evidencias sobre los procesos de respuesta utilizados en una prueba educativa estandarizada de razonamiento. Al comparar los coeficientes de concordancia calculados para el método tradicional de reportes verbales y para el RGE, se puede observar una diferencia considerable a favor de este nuevo método en cuanto al nivel de concordancia alcanzado entre los jueces.

Es importante destacar que la concordancia entre jueces es de vital importancia para la recopilación de evidencias sobre los procesos de respuesta empleados por los examinados en las pruebas educativas. Dicha importancia se basa en la necesidad de evidenciar un mínimo de objetividad a la hora de establecer si un determinado ítem mide

una determinada estrategia de resolución. Así pues, si un grupo de expertos puede alcanzar cierto consenso con respecto a lo que mide un ítem o a si un examinado empleó o no una determinada estrategia de resolución, es válido concluir que existe un respaldo aceptable sobre lo que dicho ítem evalúa. Evidentemente, tal y como lo plantean los estándares de la AERA, la APA y el NCME (2014), ningún tipo de evidencia de validez es suficiente por sí misma, sino que es necesaria la convergencia de varios tipos para alcanzar una mayor certidumbre sobre lo que se evalúa con una prueba educativa.

Por otra parte, fue posible reducir la influencia de las diferencias individuales entre los participantes en cuanto a su habilidad para reportar en voz alta sus estrategias de resolución. Este aspecto fue una ventaja comparativa importante del método RGE, ya que permite que los participantes se concentren solamente en desarrollar una estrategia de resolución de los ítems, sin la necesidad de sobrecargar sus recursos cognitivos con la tarea adicional de reportar en voz alta sus propias estrategias de resolución. De esta manera, es relevante recalcar que tanto las pruebas educativas como los métodos empleados para validar las inferencias hechas a partir de estas deben reducir al mínimo las fuentes de variabilidad irrelevantes a las habilidades que se desea medir. Así pues, aunque el método tradicional de los reportes verbales se emplea para explorar las estrategias de resolución empleadas por los examinados ante un determinado ítem, lamentablemente introduce la exigencia de que los participantes sean capaces de verbalizar con un mínimo de claridad todas las ideas y estrategias que emplean cuando resuelven un ítem.

En futuras investigaciones será necesario explorar la posible influencia de factores, como la habilidad de razonamiento en contextos verbales de los participantes y la dificultad de los ítems en los resultados obtenidos mediante el RGE. En este sentido, será relevante poner a prueba la hipótesis de que contestar correctamente los ítems depende fundamentalmente de la guía aportada por el experto en forma de pregunta y no únicamente de factores como la habilidad del participante o la dificultad del ítem. Para ello, será necesario utilizar instrumentos de medición que permitan cuantificar habilidades de gran impacto en la resolución de ítems de razonamiento, como lo son la capacidad de memoria de trabajo, las habilidades de inteligencia fluida, el bagaje léxico de los participantes, entre otros. Al ser este un estudio inicial sobre las ventajas del método RGE, no era pertinente introducir estas variables de control, ya que no se conocía *a priori* qué variables podrían afectar la resolución de los ítems en el contexto de las preguntas guía de un experto.

Finalmente, es relevante enfatizar que este nuevo método no es un sustituto del método tradicional de los reportes verbales, sino un complemento que puede ser de gran utilidad en ciertos contextos de medición y evaluación educativa. Por lo tanto, la combinación de ambos métodos es una alternativa muy útil para obtener información adicional sobre cómo se enfrentan los examinados a los ítems.

6. Referencias

American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Estados Unidos: American Educational Research Association.

Basterra, María del Rosario, Trumbull, Elise y Solano-Flores, Guillermo. (2011). Preface. En María del Rosario Basterra, Elise Trumbull y Guillermo Solano Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 12-15). Estados Unidos: Routledge.

Brizuela, Armel y Montero, Eiliana. (2013). Predicción del nivel de dificultad en una prueba estandarizada de comprensión de lectura: aportes desde la psicometría y la psicología cognitiva. *Relieve*, *19*(2), 1-23.

Brizuela, Armel, Cerdas, Danny, Fallas, Selene, Ordóñez, Kenner, Pérez, Nelson, Rojas, Luis y Seas, Guido. (2015). *Resolvamos la PAA*. Costa Rica: Editorial Universidad de Costa Rica.

Brizuela, Armel, Jiménez, Karol, Pérez, Nelson y Rojas, Guaner. (2016). Autorreportes verbales en voz alta para la identificación de procesos de razonamiento en pruebas estandarizadas. *Revista Costarricense de Psicología*, *35*(1), 17-30.

Castillo, Miguel y Padilla, José. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social indicators research*, *114*(3), 963-975.

Cromley, Jennifer y Azevedo, Roger. (2006). Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning*, *1*(3), 229-247.

Cui, Ying y Roduta, Mary. (2013). Validating Student Score Inferences with Person-Fit Statistic and Verbal Reports: A Person-Fit Study for Cognitive Diagnostic Assessment. *Educational Measurement: Issues and Practice*, *32*(1), 34-42.

Ercikan, Kadriye, Arim, Rubab, Law, Danielle, Domene, Jose, Gagnon, France y Lacroix, Serge. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24-35.

Farr, Roger, Pritchard, Robert y Smitten, Brian. (1990). A Description of What Happens When an Examinee Takes a Multiple-Choice Reading Comprehension Test. *Journal of Educational Measurement*, *27*(3), 209-226.

Fox, Mark, Ericsson, K. Anders y Best, Ryan. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological bulletin*, *137*(2), 316-344.

Garrison, Anne y Andrews-Larson, Christine. (2016). Why Don't Teachers Understand our Questions? Reconceptualizing Teachers' "Misinterpretation" of Survey Items. *AERA Open*, 2(2), 1-13. Doi: 10.1177/2332858416643077

Gorin, Joanna. (2006). Test design with cognition in mind. *Educational measurement: Issues and practice*, 25(4), 21-35.

Jiménez, Karol y Morales, Evelyn. (2009-2010). Validez predictiva del Promedio de Admisión de la Universidad de Costa Rica y sus componentes. *Actualidades en Psicología*, 23-24, 21-55.

Karabenick, Stuart, Woolley, Michael, Friedel, Jeanne, Ammon, Bridget, Blazevski, Julianne, Rhee Christina, de Groot, Elizabeth, Gilbert, Melissa, Musu, Lauren, Kempler, Toni y Kelly, Kristin. (2007). Cognitive Processing of Self-Report Items in Educational Research: Do They Think What We Mean? *Educational Psychologist*, *42*(3), 139-151. Doi https://doi.org/10.1080/00461520701416231

Landis, J. Richard y Koch, Gary. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Leighton, Jaqueline y Gierl, Mark. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3-16.

Leighton, Jaqueline y Gokiert, Rebecca. (2008). Identifying potential test item misalignment using student verbal reports. *Educational Assessment*, *13*(4), 215-242. Doi https://doi.org/10.1080/10627190802602384

Leighton, Jaqueline. (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.

Miller, Kristen, Willson, Stephanie, Chepp, Valerie y Padilla, José (Eds.). (2014). *Cognitive interviewing methodology*. Estados Unidos: John Wiley & Sons.

Padilla, José y Benítez, Isabel. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136-144.

Powers, Donald y Leung, Susan W. (1995). Answering the New SAT Reading Comprehension Questions without the Passages. *Journal of Educational Measurement*, *32*(2), 105-129.

Rojas, Luis. (2013). Validez predictiva de los componentes del promedio de admisión a la Universidad de Costa Rica utilizando el género y el tipo de colegio como variables control. *Actualidades Investigativas en Educación*, *13*(1), 1-24. Doi https://doi.org/10.15517/aie.v13i1.11707

Rojas, Luis. (2014). Evidencias de validez de la Prueba de Aptitud Académica de la Universidad de Costa Rica basadas en su estructura interna. *Actualidades en Psicología, 28*(116), 15-26. Doi https://doi.org/10.15517/ap.v28i116.14889

Rupp, André, Ferne, Tracy y Choi, Hyeran. (2006). How Assessing Reading Comprehension with Multiple-Choice Questions Shapes the Construct: A Cognitive Processing Perspective. *Language Testing*, *23*(4), 441-474.

Ryan, Katherine, Gannon-Slater, Nora y Culbertson, Michael. (2012). Improving Survey Methods with Cognitive Interviews in Small- and Medium-Scale Evaluations. *American Journal of Evaluation*, *33*(3), 414-430. Doi 10.1177/1098214012441499

Segura, Mario. (2009). La evaluación de los aprendizajes basada en el desempeño por competencias. *Actualidades Investigativas en Educación*, 9(2), 1-25. Doi https://doi.org/10.15517/aie.v9i2.9522

Shrout, Patrick y Fleiss, Joseph. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Sireci, Stephen, Baldwin, Peter, Martone, Andrea, Zenisky, April, Kaira, Leah, Lam, Wendy, Lewis, Christine, Han, Kyung, Deng, Nina, Delton, Jill y Hambleton, Ronald. (2008). *Massachusetts adult proficiency tests technical manual: Version*. Recuperado de http://www.umass.edu/remp/CEA_TechMan.html

Smith, Mark. (2017). Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes? *American Educational Research Journal*, *54*(6), 1256-1287. Doi 10.3102/0002831217717949

Smith, Vanessa y Molina, Mauricio. (2011). *La entrevista cognitiva*. Costa Rica: Instituto de Investigaciones Psicológicas.

Taylor, K. Lynn y Dionne, Jean-Paul. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, *92*(3), 413-425.

UNESCO. (2015). *Educación 2030: Declaración de Incheon y Marco de Acción para la realización del Objetivo de Desarrollo Sostenible 4*. Recuperado de http://unesdoc.unesco.org/images/0024/002456/245656s.pdf

Wagner, Daniel. (2011). *Smaller, Quicker, Cheaper: Improving Learning Assessments for Developing Countries*. Recuperado de http://unesdoc.unesco.org/images/0021/002136/213663e.pdf

Wang, Changjiang y Gierl, Mark. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, *48*(2), 165-187.

Willis, Gordon. (2015). *Analysis of the cognitive interview in questionnaire design*. Estados Unidos: Oxford University Press.

Zapata, Gerardo y Canet, Teresa. (2008). Propuesta metodológica para la construcción de escalas de medición a partir de una aplicación empírica. *Actualidades Investigativas en Educación*, 8(2), 1-26. Doi https://doi.org/10.15517/aie.v8i2.9342