Documento descargado de http://www.elsevier.es el 28-01-2017

Inv Ed Med. 2017;6(21):8–15

Investigación en Educación Médica

http://riem.facmed.unam.mx

ARTÍCULO ORIGINAL

Evaluación de reactivos de opción múltiple en medicina. Evidencia de validez de un instrumento

Jesús Rivera Jiménez[a,*], **Fernando Flores Hernández**[b], **Amilcar Alpuche Hernández**[b] y **Adrián Martínez González**[b]

[a] *Departamento de Bioquímica, Facultad de Medicina, Universidad Nacional Autónoma de México (UNAM), Ciudad de México, México*

[b] *Secretaría de Educación Médica, Facultad de Medicina, Universidad Nacional Autónoma de México (UNAM), Ciudad de México, México*

Recibido el 13 de enero de 2016; aceptado el 27 de abril de 2016
Disponible en Internet el 4 de junio de 2016

PALABRAS CLAVE
Validez de pruebas;
Reactivos de opción múltiple;
Evaluación

Resumen

Introducción: La adecuada elaboración de los reactivos de un examen constituye una evidencia de validez del mismo. A pesar de existir un consenso general sobre las recomendaciones en la elaboración de un buen reactivo, hay diferentes estudios publicados que reportan una alta incidencia de fallas en el apego a las mismas. Se propone un instrumento para evaluar la calidad en la elaboración de reactivos de opción múltiple y se describe el proceso de obtención de evidencias de validez.

Método: Se obtuvo evidencia de validez de un instrumento diseñado ex profeso para evaluar las características de los reactivos de opción múltiple, de acuerdo con las fuentes propuestas por los *Standards for Educational and Psychological Testing*, atendiendo a aquellas fuentes relacionadas con el contenido, el proceso de respuesta y la estructura interna. Se calculó el índice Kappa (por el modelo propuesto por Fleiss) y la correlación punto-biserial de Pearson para medir la concordancia en los diferentes criterios que evalúa el instrumento. Se realizó un análisis factorial exploratorio para identificar las dimensiones del instrumento y se calculó el alfa de Cronbach como estadístico de consistencia interna.

Resultados: La concordancia entre múltiples jueces tuvo un valor mayor de 0.8 (acuerdo casi perfecto) para 12 de los 21 criterios, y de 0.19 para el nivel taxonómico. El análisis factorial definió 4 dimensiones con un KMO = 0.666, ($p < 0.01$), una varianza total explicada de 49.979%, y un α de Cronbach de 0.627.

* Autor para correspondencia. Departamento de Bioquímica, Facultad de Medicina, Circuito Interior, Ciudad Universitaria, Edificio D, 1er. Piso. Av. Universidad 3000, C.P. 04510, Ciudad de México, México, Teléfono: (+5255) 5623 2300, ext. 32170.
Correo electrónico: marvin@bq.unam.mx (J. Rivera Jiménez).
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

http://dx.doi.org/10.1016/j.riem.2016.04.005
2007-5057/© 2016 Universidad Nacional Autónoma de México, Facultad de Medicina. Este es un artículo Open Access bajo la licencia CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Documento descargado de http://www.elsevier.es el 28-01-2017

Conclusión: Este instrumento puede ser aplicado para la evaluación de reactivos de opción múltiple, ya que cuenta con evidencia de validez relacionada con el contenido, el proceso de respuesta y estructura interna y los indicadores psicométricos son adecuados para su instrumentación.
© 2016 Universidad Nacional Autónoma de México, Facultad de Medicina. Este es un artículo Open Access bajo la licencia CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

KEYWORDS
Test validity;
Multiple choice ítems;
Assessment

Assessment of multiple-choice questions in medicine. Validity evidence of an instrument

Abstract
Introduction: The appropriate preparation of test ítems of an examination constitutes validity evidence in itself. Despite there being a general consensus about item-writing guidelines, several studies report a high incidence of violations of these standards. An instrument is proposed in order to assess the quality of multiple-choice item-writing, describing the validity evidence gathering process.
Methods: The validity evidence was gathered on an instrument designed to assess multiple choice ítems features, according to the sources proposed by the *Standards for Educational and Psychological Testing*, and particularly those related to content, response process, and internal structure. Kappa index (following Fleiss' model) and point-biserial correlation coefficient were used to measure concordance in the criteria assessed by the instrument. An exploratory factorial analysis was performed to identify the instrument dimensions, and Cronbach's alpha was calculated as an internal consistency statistic.
Results: Concordance between multiple judges was greater than 0.8 (almost perfect agreement) for 12 out of 21 criteria, and 0.19 for Bloom's taxonomy level. Factorial analysis defined 4 dimensions with Kaiser-Meyer-Olkin (KMO) test =0.666 (p<.01), explained variance of 49.979%, and a Cronbach's alpha of 0.627.
Conclusion: This instrument can be used to assess multiple choice ítems, since it counts with validity evidence related to content, response process and internal structure, and psychometric values appropriated for instrumentation.
© 2016 Universidad Nacional Autónoma de México, Facultad de Medicina. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introducción

La evaluación del aprendizaje es un proceso fundamental durante la formación de cualquier profesionista, ya que permite valorar el logro de los objetivos de aprendizaje por parte de los estudiantes, o de acuerdo con los modelos educativos más recientes, el logro de las competencias necesarias para el ejercicio profesional. La elección del instrumento de evaluación más adecuado depende del nivel de desempeño que se pretende evaluar, que puede ir desde el dominio cognitivo (por ejemplo, los niveles de memoria, comprensión y aplicación del conocimiento de acuerdo con la taxonomía de Bloom[1]) o el conocimiento necesario para llevar a cabo las funciones profesionales (de acuerdo con el nivel inferior de la pirámide de Miller[2]). Los exámenes con reactivos de opción múltiple (ROM) son uno de los instrumentos más ampliamente utilizados para estos fines[3]. En la literatura existen diversas revisiones sobre las características que debe de tener un ROM, una de las más citadas fue realizada por Downing, Haladyna y Rodríguez[4], quienes proponen una serie de 31 recomendaciones de acuerdo con la evidencia publicada, clasificándolas en aspectos relacionados con el contenido, el formato, el estilo, la redacción del tallo y la redacción de las opciones de respuesta, aportando evidencia de validez para el instrumento que los contiene.

El actual concepto de validez, de acuerdo con los *Standards for Educational and Psychological Testing* (un referente internacional en evaluación educativa), se refiere al grado en el que la evidencia y la teoría apoyan las interpretaciones de los resultados de una prueba para los usos propuestos del examen[5]; Downing la define como la evidencia que permite apoyar o refutar el significado o la interpretación de los resultados de una evaluación[6]. Ambas propuestas describen cinco diferentes fuentes de evidencia de validez del proceso de evaluación, relacionadas con el contenido de la prueba, el proceso de respuesta, la estructura interna, su relación con otras variables y las consecuencias para la persona que es objeto de la evaluación. Dentro de las evidencias de validez relacionadas con el contenido, un aspecto importante es la calidad de las preguntas que forman parte de la evaluación.

Diversos estudios reportan una prevalencia en las ciencias de la salud de reactivos con errores en su elaboración de hasta 50%[7–10]. Algunos estudios han reportado también que estos defectos en la elaboración tienen una repercusión en las características psicométricas de los reactivos[10,11],

Documento descargado de http://www.elsevier.es el 28-01-2017

lo cual puede tener efectos negativos en la calificación de los estudiantes[12]. Identificar estos errores en las características cualitativas (relacionadas con la calidad en su elaboración), permitirá someterlos a revisión antes de ser aplicados, mejorando la construcción del instrumento. En la literatura existen diversas propuestas de recomendaciones o directrices para la elaboración de reactivos de opción múltiple, las cuales son muchas veces versiones resumidas y adaptadas de la propuesta de Haladyna para las necesidades de alguna institución; en nuestra búsqueda encontramos únicamente algunas listas de cotejo en idioma inglés para la evaluación de reactivos de opción múltiple, pero sin estudios que aporten evidencia de validez para su uso[13,14], por lo que es necesario tener un instrumento que cuente con esta evidencia, para poder realizar una evaluación del aprendizaje más objetiva.

El objetivo de este artículo es proponer un instrumento en español que permite evaluar las características cualitativas de un reactivo de opción múltiple, de acuerdo con las recomendaciones propuestas en la literatura y se describe el proceso de obtención de evidencia de validez.

Método

Se realizó un estudio no experimental de tipo descriptivo para la elaboración del instrumento, obteniendo la evidencia de validez del uso del mismo en cada etapa de la metodología. Se realizó en la Facultad de Medicina de la UNAM, utilizando 308 reactivos de opción múltiple aplicados en exámenes parciales sumativos durante los años escolares 2012, 2013 y 2014.

Procedimiento

Etapa 1. Evidencia de validez relacionada con el contenido

Se tomó como modelo la propuesta de Haladyna, Downing y Rodríguez[4] para elaborar un instrumento, a manera de lista de cotejo, que permitiera evaluar la presencia de las características deseables para un buen reactivo de opción múltiple. A partir de dos diferentes traducciones de las recomendaciones de Haladyna[15,16], se elaboró una tabla comparativa para elegir aquellas recomendaciones que fueran más claras y precisas, las cuales fueron convertidas en afirmaciones y pasaron a formar parte del instrumento, generando una propuesta inicial de 24 ítems. Se decidió utilizar las definiciones propuestas por Krathwohl[1] y de Buckwalter[17] para los niveles de la taxonomía de Bloom. Este primer instrumento se sometió a un proceso de validación de contenido por jueces (cinco profesores con estudios de posgrado en el área de educación y con experiencia laboral en evaluación educativa y elaboración de reactivos) contando para ello con un aula virtual en la plataforma Moodle (versión 2.6) utilizando la herramienta «lista de cotejo»; se seleccionaron 10 reactivos piloto de diferentes asignaturas de la carrera de medicina para este proceso. Al final se solicitó a los evaluadores sus recomendaciones en relación con la claridad de los criterios, aspectos que pudieran ser incluidos, así como criterios que pudieran omitirse o combinarse. Estas recomendaciones se revisaron en conjunto con otras propuestas para elaborar reactivos[15,18,19], lo que permitió reestructurar el instrumento para continuar con el proceso de validación.

Etapa 2. Evidencia de validez relacionada con el proceso de respuesta

El instrumento resultado de la validación por expertos, conformado por 22 criterios, fue aplicado para evaluar la calidad de reactivos de opción múltiple de exámenes sumativos de una asignatura biomédica de la carrera de medicina. Como criterio de inclusión para los reactivos utilizados se consideraron aquellos con formato convencional (un tallo o enunciado en donde se plantea la pregunta y una serie de opciones de respuesta en donde solo una es correcta)[7], se excluyeron aquellos reactivos con formato de verdadero/falso y de emparejamiento. El instrumento fue integrado en un aula virtual en Moodle versión 2.8, con la herramienta de «retroalimentación». La evaluación la realizaron un total de 9 profesores de la asignatura (5 con estudios de posgrado y 4 con licenciatura), ocho de ellos distribuidos aleatoriamente en parejas, mientras que un profesor participó en la evaluación de todos los reactivos (como criterio de desempate). Se capacitó a los profesores previamente en el uso y llenado del instrumento, se discutieron las dudas que surgieron al respecto del significado de algunos criterios y del uso adecuado de la taxonomía de Bloom. Los reactivos fueron distribuidos de manera aleatoria entre las parejas de profesores. Se decidió utilizar el estadístico kappa de Fleiss[20] para medir el grado de acuerdo entre múltiples jueces para cada criterio que contempla el instrumento.

El análisis estadístico para los índices kappa se realizó con la versión 2013 de Microsoft Excel®. La participación de los profesores en la prueba piloto y en la aplicación del instrumento fue totalmente voluntaria.

Para esta investigación se tomó en cuenta y no se violó ningún principio de la Declaración de Helsinki, principalmente porque el objeto de medición son los reactivos y no los estudiantes ni los profesores. La participación de los diferentes evaluadores a lo largo del estudio fue voluntaria.

Etapa 3. Evidencia de validez relacionada con la estructura interna

Las fuentes de evidencia de validez relacionadas con la estructura interna del instrumento incluyen aspectos de análisis de discriminación de ítems, confiabilidad y la estructura factorial. Para el primero, se utilizaron la correlación punto-biserial de Pearson (Rpbis) y una prueba t de Student, para la confiabilidad se utilizó el alfa de Cronbach como un estadístico de consistencia interna, y un análisis factorial exploratorio para definir la estructura final del instrumento. Estos análisis se realizaron con la versión 21 del software SPSS®.

Resultados

Evidencia de validez relacionada con el contenido

Las aportaciones del grupo de revisores expertos permitieron la generación del instrumento final, a partir de las

Documento descargado de http://www.elsevier.es el 28-01-2017

Tabla 1 Concordancia interjueces en la propuesta de instrumento

Criterio		Índice kappa
	Nivel taxonómico (memoria, comprensión, aplicación)	0.1899
1	¿El reactivo presenta un solo contenido temático?	0.7955
2	¿El reactivo presenta un solo resultado de aprendizaje?	0.7244
3	¿El contenido evaluado está en relación con la especificación del reactivo?	0.9003
4	¿El contenido del reactivo se refiere a una evidencia y no a una opinión?	0.9853
5	¿La semántica utilizada está de acuerdo con el contenido del programa académico?	0.9457
6	¿Las opciones de respuesta se presentan en vertical?	0.9853
7	¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?	0.5906
8	¿La cantidad de texto en el tallo es adecuada para su comprensión?	0.7739
9	¿El tallo del reactivo plantea la idea central?	0.8643
10	¿La pregunta o instrucción se encuentra redactada con claridad?	0.5724
11	¿Es posible responder la pregunta sin necesidad de observar las respuestas?	-0.2318
12	¿El reactivo está expresado en forma positiva (es decir, no incluye palabras como NO o EXCEPTO)?	0.9705
13	¿El reactivo cuenta con tres o cuatro opciones de respuesta?	0.8276
14	¿El reactivo cuenta únicamente con una respuesta correcta?	0.8798
15	¿Las opciones son independientes entre sí?	0.9105
16	¿Las opciones son similares en cuanto a estructura gramatical, contenido y extensión?	0.6086
17	¿Las opciones se expresan de manera afirmativa?	0.9755
18	¿Los distractores son plausibles, es decir, no se descartan por inferencia lógica o sentido común?	0.6558
19	¿Las opciones evitan dar pistas sobre la respuesta correcta?	0.7575
20	¿Se evita el uso de términos como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?	1
21	¿Se evita el uso de las opciones «Todas las anteriores» o «Ninguna de las anteriores»?	0.9902

Los valores kappa obtenidos para la concordancia entre jueces en la aplicación de los diferentes criterios del instrumento indican que 18 de los 21 ítems analizados obtienen puntajes apropiados (de 0.060 a 1).

diferentes recomendaciones sobre la redacción de los criterios, la unificación de algunos de ellos, y la inclusión de la clasificación taxonómica del reactivo como un indicador de la calidad del mismo. El nuevo instrumento quedó conformado por 21 criterios o ítems, más la clasificación del reactivo a evaluar de acuerdo con la taxonomía modificada de Bloom (tabla 1).

Evidencia de validez relacionada con el proceso de respuesta

Con el instrumento resultado del proceso de validación por expertos fueron evaluados un total de 308 reactivos de opción múltiple convencionales.

El índice kappa para el nivel taxonómico fue de 0.19, lo cual nos habla de poco acuerdo entre los jueces[21]. Para 20 de los 21 criterios de calidad el índice kappa fue desde 0.57 hasta 1, lo cual nos habla de un acuerdo desde moderado hasta perfecto (tabla 1). El criterio 20 se comportó como una constante, por lo que no se incluyó en el resto de los análisis. El criterio 11 («¿Es posible responder la pregunta sin necesidad de observar las respuestas?») obtuvo un índice kappa negativo (-0.23).

Evidencia de validez relacionada con estructura interna

La mayoría de las correlaciones entre los criterios reflejaron poca fuerza de asociación (menores de 0.1 y mayores de -0.1), aunque hubo 5 correlaciones entre ítems mayores de 0.40 ($p < 0.01$) (tabla 2). El análisis de confiabilidad resultó en un alfa de 0.615 para los 20 elementos cuantificados; al eliminar el criterio 11, la confiabilidad subió hasta 0.641. La prueba T determinó una $p > 0.01$ para los criterios 4, 6, 12, 17 y 21, por lo que se decidió excluirlos de la versión final del instrumento. Con los 14 criterios restantes se realizó el análisis factorial exploratorio, realizando una rotación oblicua por el método Oblimin (tabla 3). La media de adecuación muestral de Kaiser-Meyer-Olkin fue de 0.666 y la prueba de esfericidad de Bartlett arrojó un [$\chi^2 = 599.285$, $r < 0.01$]. A partir de este análisis se identificaron cinco factores, cuatro con al menos tres indicadores en su estructura y uno con solo un indicador. La varianza explicada con 4 factores fue de 49.979 y con los 5 fue de 57.561. El alfa de Cronbach para el instrumento con los 14 ítems restantes fue de 0.627.

Discusión

Al ser este cuestionario un instrumento de evaluación para las características de los reactivos de opción múltiple, las fuentes de evidencia de validez corresponden a las mencionadas por los *Standards*[5] de la *American Educational Research Association* y Downing[6]; el estudio aporta evidencias de validez en relación con el contenido, el proceso de respuesta y la estructura interna del mismo. La representatividad del dominio (definido como las características de un reactivo bien elaborado) está en relación con la propuesta de Haladyna, mientras que la calidad de los ítems que conforman al instrumento fue lograda al ser revisada por diversos jueces con experiencia en el área de la evaluación educativa (lo cual, a su vez, es una fuente de evidencia

Documento descargado de http://www.elsevier.es el 28-01-2017

Tabla 2 Correlaciones entre criterios del instrumento

Criterio A	Criterio B	Correlación
1 ¿El reactivo presenta un solo contenido temático?	2 ¿El reactivo presenta un solo resultado de aprendizaje?	0.704 ($p < 0.01$)
8 ¿La cantidad de texto en el tallo es adecuada para su comprensión?	10 ¿La pregunta o instrucción se encuentra redactada con claridad?	0.521 ($p < 0.01$)
8 ¿La cantidad de texto en el tallo es adecuada para su comprensión?	9 ¿El tallo del reactivo plantea la idea central?	0.497 ($p < 0.01$)
14 ¿El reactivo cuenta únicamente con una respuesta correcta?	15 ¿Las opciones son independientes entre sí?	0.424 ($p < 0.01$)
9 ¿El tallo del reactivo plantea la idea central?	10 ¿La pregunta o instrucción se encuentra redactada con claridad?	0.417 ($p < 0.01$)
7 ¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?	10 ¿La pregunta o instrucción se encuentra redactada con claridad?	0.415 ($p < 0.01$)

Las diferentes correlaciones entre los criterios demuestran que el criterio 10 correlaciona de manera significativa directa y moderada con los reactivos 7, 8 y 9; de la misma forma, el reactivo 8 correlaciona con los reactivos 9 y 10. En las correlaciones obtenidas se aprecia además congruencia teórica entre los reactivos.

de validez por sí misma); sus recomendaciones y posterior discusión permitieron elaborar un instrumento que fuera entendible y fácilmente aplicable para los evaluadores de reactivos.

La baja concordancia (kappa de 0.19) obtenida en este estudio para la asignación del nivel cognitivo que evalúa un reactivo ha sido demostrada previamente[22,23], ya que esta clasificación depende en gran medida del nivel educativo de la persona que se enfrenta al mismo: una conducta mental para un individuo puede ser diferente para otro (un estudiante puede utilizar el juicio clínico para asignar un tratamiento a una enfermedad, mientras que otro puede reconocer esta asociación solo por haberlo memorizado), aunque existen reportes que afirman que se logra un mayor acuerdo en la categorización de los niveles al involucrar a los docentes que imparten la asignatura[24].

Tabla 3 Matriz de estructura de análisis de componentes principales

Factor	Criterio del instrumento final	Carga	Alfa de Cronbach
Comprensión del reactivo	1. ¿La cantidad de texto en el tallo es adecuada para su comprensión?	0.738	0.668 Sig 0.000
	2. ¿La pregunta o instrucción se encuentra redactada con claridad?	0.721	
	3. ¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?	0.656	
	4. ¿El tallo del reactivo plantea la idea central?	0.655	
Contenido del reactivo	5. ¿El reactivo presenta un solo resultado de aprendizaje?	0.865	0.615
	6. ¿El reactivo presenta un solo contenido temático?	0.849	Sig 0.000
	7. ¿La semántica utilizada está de acuerdo con el contenido del programa académico?	0.411	
Precisión del reactivo	8. ¿El reactivo cuenta únicamente con una respuesta correcta?	0.814	0.477 Sig 0.230
	9. ¿Las opciones son independientes entre sí?	0.783	
	10. ¿El contenido evaluado está en relación con la especificación del reactivo?	0.335	
Redacción de opciones de respuesta	11. ¿Las opciones son similares en cuanto a estructura gramatical, contenido y extensión?	0.711	0.357 Sig 0.44
	12. ¿Las opciones evitan dar pistas sobre la respuesta correcta?	0.642	
	13. ¿Los distractores son plausibles, es decir, no se descartan por inferencia lógica o sentido común?	0.608	
*	14. ¿El reactivo cuenta con tres o cuatro opciones de respuesta?	0.773	-

El análisis de factorial tuvo un índice KMO de 0.666, sig Bartlett 0.000, con 7 iteraciones. El α de Cronbach para el total del instrumento fue de 0.627.

* Se consideró como indicador, ya que no hubo más criterios con los cuales conformar un factor.

Documento descargado de http://www.elsevier.es el 28-01-2017

La alta fiabilidad en las respuestas de los jueces al aplicar los distintos criterios permiten identificar que el instrumento es claro y puede ser aplicado de manera adecuada por diferentes jueces. La concordancia para el criterio 20 fue de 1 (acuerdo perfecto), dado que este es uno de los indicadores que no está sujeto a la interpretación del evaluador (al igual que otros, cuyo nivel de concordancia estuvo por arriba de 0.8, como son los ítems 6, 12, 13, 17 y 21); el resto de los criterios tienen cierta subjetividad, ya que no pretenden evaluar hechos absolutos, sino la apreciación del evaluador hacia determinadas características. Estos indicadores no mostraron tanta variabilidad en la muestra de reactivos en la que fueron aplicados, por lo que para nuestra población de reactivos en la que fue realizado el estudio no aportaron información que permitiera discriminar a los reactivos de acuerdo con su calidad. Esto no implica necesariamente que estas recomendaciones deban de omitirse durante la elaboración o evaluación de reactivos y como se ha mencionado previamente[4], pueden representar puntos de «buena práctica»

La correlación entre los distintos criterios nos permite identificar una asociación entre los ítems que conforman el instrumento. Encontramos correlaciones altas para algunos ítems; tres de ellos se asocian con el criterio 10 del instrumento («¿La pregunta o instrucción se encuentra redactada con claridad?»), el 7 se refiere a la correcta gramática, puntuación y ortografía del reactivo, el 8 a la presencia de una extensión adecuada del tallo y la 9 a la existencia de un tallo enfocado (que contenga la idea central), es sencillo identificar que estos aspectos favorecen la adecuada lectura del reactivo y por lo tanto permiten calificarlo como claro en cuanto a su redacción; la asociación entre los criterios 14 y 15 sugiere que la existencia de opciones independientes entre sí (una no incluye parcial o totalmente a otra) disminuye la probabilidad de que el reactivo pueda tener dos respuestas correctas. Finalmente, es probable que un único resultado de aprendizaje (mencionado como *mental behavior* en las recomendaciones de Haladyna) esté asociado únicamente a un contenido temático en una pregunta (incluso pretender evaluar varios contenidos temáticos en un reactivo resulta difícil al momento de elaborarlo). Dado que el ítem 20 («¿Se evita el uso de términos como SIEMPRE, NUNCA, COMPLETAMENTE o ABSOLUTAMENTE?») tuvo un grado de acuerdo total por todos los jueces, no se pudo hacer la correlación con los demás criterios, ya que se comportó como una constante.

La prueba T permitió identificar 5 ítems (4, 6, 12, 17 y 21) que no permitían discriminar las características de los reactivos evaluados, y considerando que no aportan información suficiente, fueron eliminados de la versión final del instrumento. Para eliminar el ítem 11 se tomaron en cuenta los siguientes aspectos:

- Es el único criterio del instrumento que no está incluido en el marco de elaboración de reactivos de Haladyna.
- La concordancia entre jueces (con el índice kappa) tuvo un valor negativo, lo que representa un valor menor al esperado por el azar, lo que demuestra que fue un reactivo poco confiable en su aplicación.
- La confiabilidad del instrumento aumenta al eliminar el ítem (de 0.615 a 0.641).
- El reactivo no aporta información de contenido en la estructura de la prueba; aunque su peso factorial es alto, no se asocia a ningún factor.

El análisis factorial identificó cuatro factores diferentes y un indicador, de los cuales los primeros cuatro logran explicar casi el 50% de la varianza total. El primer factor (al cual se le denominó «comprensión del reactivo») está conformado por los ítems 8, 10, 9 y 7, los mismos que habían demostrado una alta correlación; este factor hace referencia a la adecuada redacción del reactivo para permitir su comprensión. El segundo factor permitió conformar el componente «contenido del reactivo», incorporando a los ítems 2, 1 y 5; este hace referencia a la pertinencia del contenido del reactivo con lo propuesto por el programa académico. El tercer factor está integrado por los ítems 14, 15 y 3, y de acuerdo con estos criterios generó el componente «precisión del reactivo»; este factor se refiere a la relación del reactivo con una sola especificación y una única respuesta correcta. Finalmente, el cuarto factor lo conforman los ítems 16, 19 y 18, denominado «redacción de opciones de respuesta», que hacen referencia a la adecuada redacción de estos componentes. El ítem 13 tuvo una carga factorial alta, pero se decidió mantenerlo como un indicador individual, ya que representa un punto importante en la elaboración de reactivos de opción múltiple convencionales[4], a pesar de no existir reactivos suficientes para integrar un factor.

Estos componentes permitieron integrar la versión final del instrumento de evaluación.

Las recomendaciones propuestas por Haladyna[4] son indicadores sobre la evidencia de validez relacionada con el contenido para elaborar ROM. Más recientemente, Moreno et al.[25] realizaron una propuesta para elaborar ROM, aportando evidencia de validez relacionada con el contenido[15]. Ninguna de estas hace referencia a las otras fuentes de evidencia de validez reportadas en los *Standards*[5], por lo que esta investigación aporta, aparte de evidencias de validez de contenido, evidencias relacionadas con el proceso de respuesta y la estructura interna del instrumento, lo cual apoya el uso del mismo para el fin que fue diseñado.

La presencia de cualquier falla en la elaboración de un reactivo de opción múltiple nos debe llevar a revisar el reactivo y modificarlo, ya que algunas violaciones a estas recomendaciones pueden modificar el comportamiento psicométrico del ítem[10]; para otras no existe evidencia de esto, por lo que posiblemente el apegarse a ellas representa solo una «buena práctica», como se mencionó previamente; por esto es necesario realizar más estudios para determinar la pertinencia de todos los criterios incluidos en este instrumento. Haladyna[4] reporta que algunas recomendaciones son citadas en más del 90% de las guías consultadas en su última revisión publicada sobre el tema.

En el instrumento se mencionan algunos términos como «especificación», «resultado de aprendizaje», «tallo», «distractores», los cuales pudieran resultar familiares solo a aquellas personas con experiencia en los procesos de evaluación. Esto asegura, de algún modo, que aquellas personas que evalúen los reactivos posean experiencia en evaluación y elaboración de reactivos, ya que una fuente de evidencia de validez es la capacitación de las personas que elaboran y

Documento descargado de http://www.elsevier.es el 28-01-2017

revisan los reactivos, esto ha demostrado mejorar la calidad de los reactivos de un examen[14].

Este estudio se realizó con reactivos de una asignatura de la licenciatura de Médico Cirujano, por lo que es necesario obtener evidencia del uso del instrumento en otras asignaturas, incluso en otras áreas de conocimiento dentro y fuera de las ciencias de la salud (fuentes de evidencia de validez relacionadas con otras variables[5]). Es importante también analizar la relevancia de cada uno de los criterios en el desempeño psicométrico del reactivo evaluado.

El impacto de contar con un instrumento de este tipo contribuirá a complementar y consolidar los procesos de las evaluaciones de altas consecuencias en la educación en ciencias de la salud.

Conclusiones

El instrumento descrito en el presente artículo permite evaluar las características cualitativas de un reactivo de opción múltiple, de acuerdo con las recomendaciones que se proponen en la literatura, obteniendo evidencias de validez y confiabilidad relacionadas con el contenido del instrumento, el proceso de respuesta y la estructura interna. Es importante que los evaluadores también lleven a cabo un proceso de capacitación para la elaboración y revisión de reactivos, para que la aplicación del instrumento como estrategia de evaluación de los reactivos sea, a su vez, válida, confiable y objetiva.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Financiación

Ninguna.

Autoría/colaboradores

JRJ, FFH y AMG participaron en el proceso de elaboración del instrumento. JRJ y AAH participaron en el análisis de los resultados. Todos los autores participaron en la discusión, la elaboración del manuscrito y su revisión.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Agradecimientos

Los autores agradecen al doctor Iwin Leenen por el apoyo en el análisis estadístico, a los profesores de la Secretaría de Educación Médica que participaron en el proceso de validación de contenido del instrumento y a los profesores que participaron en la aplicación del instrumento para la evaluación de los reactivos.

Referencias

1. Krathwohl DR. A revision of Bloom's taxonomy: An Overview. Theory Pract. 2002;41:212–8.
2. Miller GE. The assessment of clinical skills-competence-performance. Acad Med. 1990;65:S63–7.
3. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357:945–9.
4. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. 2002;15:309–34.
5. American Educational Research, Association, American Psychological Association, National Council on Measurement in Education. The standards for educational and psychological testing. Washington, D.C.:American Educational Research Association, 2014.
6. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003 Sep;37:830–7.
7. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of ítem writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Pract. 2006 Dec;6:354–63.
8. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77:156–61.
9. Masters JC, Hulsmeyer BS, Pike ME, Leichty K, Miller MT, Verst AL. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. J Nurs Educ. 2001 Jan;40:25–32.
10. Pate A, Caldwell DJ. Effects of multiple-choice item-writing guideline utilization on ítem and student performance. Curr Pharm Teach Learn. 2014 Jan;6:130–4.
11. Jurado-Nuñez AG, Flores-Fernandez F, Delgado-Maldonado L, Sommer-Cervantes H, Martínez-González A, Sánchez-Mendiola M. Distractores en preguntas de opción múltiple para estudiantes de Medicina ¿Cuál es su comportamiento en un examen de altas consecuencias? Inv Ed Med. 2013;2:202–10.
12. Downing SM. The effects of violating standard ítem writing principles on test and students: The consequences of using flawed test ítems on achievement examinations in medical education. Adv Heal Sci Educ. 2005;10:133–43.
13. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ Theory Pract. 2012 Aug;17:369–76.
14. Tarrant M, Ware J. A framework for improving the quality of multiple-choice assessments. Nurse Educ. 2012;37:98–104.
15. Moreno R, Martínez RJ. Directrices para la construcción de ítems de elección múltiple. Psicothema. 2004;16:490–7.
16. Downing SM, Haladyna TM. Manual para el desarrollo de pruebas a gran escala. México, D.F: Centro Nacional de Evaluación para la Educación Superior; 2012.
17. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an educational taxonomy for evaluation of cognitive performance. J Med Educ. 1981;56:115–21.
18. Case SM, Swanson DB. Cómo construir preguntas de selección múltiple para ciencias básicas y ciencias clínicas. Philadelphia: National Board of Medical Examiners; 2014.

Documento descargado de http://www.elsevier.es el 28-01-2017

19. Dirección General de Evaluación Educativa UNAM. Lineamientos generales para la elaboración de reactivos [Internet]. [citado 4 Abr 2015]. Disponible en: http://www.inb.unam.mx/ensenanza/lineamto_gral_elabora_reactivo.pdf.
20. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378–82.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
22. Cunnington JPW, Norman GR, Blake JM, Dauphinee WD, Blackmore DE. Applying learning taxonomies to test items: is a fact an artifact? Acad Med. 1996;71:31–3.
23. Kibble JD, Johnson T. Are faculty predictions or ítem taxonomies useful for estimating the outcome of multiple-choice examinations? AJP: Adv Physiol Educ. 2011;35:396–401.
24. Thompson E, Luxton-Reilly A, Whalley JL, Hu M, Robbins P. Bloom's taxonomy for CS assessment. Conf Res Pract Inf Technol Ser. 2008;78:155–61.
25. Moreno R, Martínez RJ, Muñiz J. New guidelines for developing multiple-choice items. Methodology. 2006;2:65–72.