TRABAJO ORIGINAL

Índice de calidad para evaluar preguntas de opción múltiple

ALBERTO GALOFRÉ T.[1] y ANA C. WRIGHT N.[2]

RESUMEN

El índice de calidad para evaluar preguntas de opción múltiple, que se describe en este trabajo, se desarrolló para indicar con una nota o puntaje la calidad relativa de este tipo de preguntas. Se basa en diez atributos reconocidos como integrantes de una pregunta de opción múltiple bien construida: presencia de viñeta (presencia de un caso clínico o procedimiento de laboratorio o problema), enunciado completo (cuando al cubrir las opciones el lector puede responder a la pregunta después de solamente leer el enunciado), evitar uso de negaciones, concordancia gramatical entre enunciado y opciones, distractores verosímiles, extensión similar entre las opciones, evitar ninguna y todas las anteriores, opciones ordenadas, opciones homogéneas (similares en cuanto al contenido) y aplicación de conocimiento o superior. El índice se calcula a partir de un análisis de los defectos presentes en cada pregunta en una escala de 1 a 5, siendo 5 el puntaje óptimo de calidad y 1 el menor, en base al siguiente detalle: 5 = pregunta con viñeta, sin defectos de construcción; 4 = pregunta sin viñeta, sin defectos de construcción; 3 = pregunta con o sin viñeta, con un defecto; 2 = pregunta con o sin viñeta, con dos defectos; 1 = pregunta con o sin viñeta, con tres o más defectos. El objetivo principal de este índice es representar con un puntaje la calidad relativa de una pregunta de opción múltiple, simplificando el proceso para que sea posible implementar por docentes y educadores en ciencias de la salud con instrucciones claras respecto de su uso. El índice de calidad sirve en varios escenarios, por ejemplo: analizar la calidad de las preguntas de un examen, resumir la calidad global de un examen con el promedio del índice, comunicación con redactores de preguntas sobre la relativa calidad de cada una y cómo se puede mejorar, y además como un mecanismo de feedback formativo en talleres y sesiones para mejorar la calidad de la producción de preguntas de opción múltiple.

Palabras clave: Índice de calidad, preguntas de opción múltiple, evaluación.

SUMMARY

A quality index to evaluate multiple choice questions

The index described in this paper was developed to indicate with a score the relative quality of a multiple-choice item. It is based on ten selected guidelines recommended for well constructed multiple choice items: presence of a vignette (a clinical or laboratory procedure or problem), clear question on the stem enabling arriving at an answer with the options covered, avoid use of negative words on the stem (for example EXCEPT), grammatical consistency between the question in the stem and options, plausible distractors, similar length of the options, options ordered numerically or logically, homogeneous distractors (similar in content), and whether the item assesses application of knowledge or higher. The index is calculated from an analysis of the defects present in each item on a scale of 1 to 5, with 5 being the best quality score and 1 the lowest, based on the following scale: 5 = item with a vignette and flawless construction, 4 = item without vignette, flawless construction, 3 = item with or without a vignette, with one defect, 2 = item with or without vignette and two defects, 1 = item with or without a vignette and three or more defects. The main

Recibido: el 20/08/10, Aceptado: el 8/10/10.

[1] MD, MEd. Profesor emérito, Saint Louis University School of Medicine, St. Louis, Missouri, USA.
[2] Profesor Asistente, Centro de Educación Médica, Pontificia Universidad Católica de Chile, Chile.

objective of this index is to use a score designed to represent the relative quality of a multiple-choice item. Furthermore, the process of producing the score has been simplified for implementation by health sciences educators using a straightforward procedure. The index can be used in various scenarios, for example to assess the quality of individual items when selecting items for a test, to summarize the overall quality of a test using the mean of individual scores, to mentor item writers on the relative quality of their items and how the items can be improved, and also as a feedback mechanism in workshops and training sessions to enhance the quality of the production of multiple-choice items.

Key words: Quality index, MCQ, assessment.

INTRODUCCIÓN

El índice de calidad para evaluar preguntas de opción múltiple, que se describe en este trabajo, se desarrolló para indicar con una nota o puntaje la calidad relativa de una pregunta de opción múltiple. Se basa en diez atributos reconocidos como integrantes de una pregunta de opción múltiple bien construida.

Este índice está basado en el trabajo de Josefowicz et al.[1] Ese trabajo presentó un índice de calidad en que cada pregunta era evaluada en una escala de 5 puntos. Así, una pregunta que solamente evaluaba recuerdo de hechos aislados y además presentaba defectos de construcción, recibía un puntaje de 1; al otro extremo, una pregunta que contenía una viñeta (presencia de un caso clínico, o procedimiento de laboratorio, o problema en el enunciado de la pregunta), requería un proceso de razonamiento, y no tenía defectos de construcción, recibía un puntaje de 5.

Aunque el sistema mencionado tiene méritos, su aplicación no se ha extendido. A nuestro juicio, esto se podría deber a que las instrucciones para llegar al puntaje o nota de cada pregunta son relativamente complejas y en cierto modo vagas. Así, en su aspecto más básico, se puede decir que el procedimiento usado por Josefowicz et al. usa dos elementos fundamentales: la presencia de una viñeta y que no haya defectos técnicos de construcción de preguntas (basado principalmente en el manual de S. Case y D. Swanson, del National Board of Medical Examiners (NBME))[2]. Pero hay al menos dos docenas de defectos y recomendaciones que se mencionan en ese manual. Es difícil mantener en la memoria tantos elementos al evaluar una pregunta. Además, en las dos publicaciones[1,3] en que se ha usado este índice, el procedimiento fue implementado con expertos del NBME e incluso autores del manual indicado.

Nuestro objetivo fue tomar las ideas de Josefowicz et al. y tratar de simplificar el proceso haciéndolo más transparente y objetivo y que fuera posible implementar por docentes y educadores en ciencias de la salud con instrucciones claras respecto de su uso.

MATERIAL Y MÉTODO

Construcción del índice

Se hizo una revisión de varios textos y publicaciones para hacer un catastro de recomendaciones de calidad de preguntas de opción múltiple[2,4,5]. De ellas, se eligió diez recomendaciones que nos parecieron fueran relativamente fáciles de implementar y capaces de representar los defectos de construcción comunes.

Los factores que se tomaron en cuenta para evaluar la calidad técnica de cada pregunta son: presencia de viñeta, enunciado completo, evitar uso de negaciones, concordancia gramatical entre enunciado y opciones, distractores verosímiles, extensión similar entre las opciones, evitar ninguna y todas las anteriores, opciones ordenadas, opciones homogéneas y aplicación de conocimiento o superior.

A continuación se describe cada factor.

Presencia de viñeta: Se refiere a la presencia de un caso clínico o procedimiento de laboratorio o problema.

Enunciado completo: Cuando al cubrir las opciones el lector puede responder a la pregunta después de solamente leer el enunciado.

Evita uso de negaciones: Cuando el enunciado no pregunta por lo que no es, como Excepto, Falsa.

Concordancia gramatical entre enunciado y opciones: Que la pregunta del enunciado y las opciones sigan reglas gramaticales. Por ejemplo, si se pregunta por "el" no poner una palabra femenina después en una de las opciones.

Distractores verosímiles: Que cada opción parezca verdadera para el que no sabe mucho.

Extensión similar entre las opciones: Es usual que si existe una opción con excesiva longitud en relación a las otras, tiende a ser la respuesta correcta por tener mayor detalle.

Evitar ninguna y todas las anteriores: Bastaría saber que dos opciones son correctas para saber que si existe una opción de "todas las anteriores" ésta tiene que ser la respuesta. "Ninguna de las anteriores" se presta para interpretaciones diferentes del que redacta la pregunta. Por ejemplo, un mejor tratamiento o la mejor conducta a seguir después

de dar los antecedentes. En formato de pregunta de "una mejor respuesta" estas opciones complican el proceso de responder.

Opciones ordenadas: Ya sea lógicamente o numéricamente. Si hay números, por ejemplo, 5, 10, 15, 20, 25, estos deben colocarse en ese orden. No es recomendable poner opciones con números sin ordenar. Complica la respuesta innecesariamente.

Opciones homogéneas: Similares en cuanto al contenido.

***Aplicación de conocimiento o superior*:** Que la pregunta requiera usar un proceso superior a simple recuerdo de un hecho memorizado.

Cálculo del índice para cada pregunta

El índice se calcula a partir de un análisis de los defectos presentes en cada pregunta. El resultado se expresa en una escala de 1 a 5, siendo 5 el puntaje óptimo de calidad y 1 el menor. Al combinar la presencia de viñeta y la cantidad de defectos se logra la siguiente escala de puntajes:

5 = pregunta con viñeta, sin defectos de construcción.
4 = pregunta sin viñeta, sin defectos de construcción.
3 = pregunta con o sin viñeta, con un defecto.
2 = pregunta con o sin viñeta, con dos defectos.
1 = pregunta con o sin viñeta, con tres o más defectos.

Para facilitar el cálculo del índice para cada pregunta, se recomienda construir una planilla que incorpore columnas para cada factor con hileras para cada pregunta. Además, se agrega una columna inicial para identificar la pregunta, habitualmente con el número correspondiente, y se complementa con una última columna destinada al índice resultante, tal como aparece en la Figura 1. El procedimiento para evaluar cada pregunta es sencillo: se examina cada uno de ellos para la presencia o no de cada uno de los factores incluidos en el índice. Nos ha sido útil usar un "1" si el factor está presente y un "0" si está ausente. Para calcular el índice se examina primero la columna de la viñeta y luego los factores restantes y se aplica la escala para calcular el índice. Es decir, no se suman los "puntos" logrados, sino que se cuentan los defectos; en nuestra pauta se cuentan los números "0" (cero). En las Tablas 1 y 2 aparecen dos ejemplos de aplicación del índice.

DISCUSIÓN

La evaluación del conocimiento en ciencias de la salud se hace, en gran medida, con preguntas de opción múltiple. Por tanto, es importante que la confección de dichas preguntas sea de la mejor calidad posible. Desgraciadamente, pareciera que esto no ocurre con la frecuencia necesaria. Josefowicz et al.[1], en una investigación realizada con preguntas obtenidas de tres escuelas de medicina en los Estados Unidos, encontraron que cuando las preguntas eran redactadas por docentes sin entrenamiento en confeccionar preguntas de calidad, se lograba un índice de alrededor de sólo 2. Downing[6] analizó la presencia de preguntas con defectos en exámenes de asignaturas en medicina y encontró que las preguntas mal construidas tenían un índice de dificultad mayor que las bien construidas. Esta situación es importante, pues esta dificultad es un factor sin relación al contenido, que tiene consecuencias para aquellos alumnos con puntajes cercanos al corte del puntaje de aprobación.

El presente índice se ha estado usando desde el año 2005. En nuestra experiencia, al utilizar el índice en una muestra de exámenes en educación médica en Chile, encontramos un índice promedio de 1,6 en un examen final de asignatura. En una muestra de

	Presenta viñeta	Enunciado Completo	Evita uso de negociaciones	Concordancia gramatical entre enunciado y opciones	Distractores verosímiles	Extensión similar entre las opciones	Evita el uso de Ninguna y Todas las anteriores	Opciones ordenadas	Opciones homogéneas	Aplicación de conocimientos o superior	Puntaje

Figura 1. Modelo de la planilla usada para evaluar las preguntas y calcular el índice.

Tabla 1. Ejemplo 1 de aplicación del índice

Pregunta

Mujer de 35 años que consulta por cuadro febril de 15 días de evolución, poliartralgias, artritis de muñecas y de metacarpofalángicas. Al examen se constata eritema facial. Hemograma: leucopenia y linfopenia. Sedimento urinario: proteinuria ++. ¿Cuál es el diagnóstico más probable?

a) Dermatomiositis
b) Artritis reumatoidea
c) Lupus eritematoso sistémico
d) Esclerosis sistémica progresiva
e) Vasculitis

Índice

	Presenta viñeta	Enunciado Completo	Evita uso de negociaciones	Concordancia gramatical entre enunciado y opciones	Distractores verosímiles	Extensión similar entre las opciones	Evita el uso de Ninguna y Todas las anteriores	Opciones ordenadas	Opciones homogéneas	Aplicación de conocimientos o superior	Puntaje
	0	0	1	1	1	1	1	n/c	0	0	1

n/c = No corresponde

Tabla 2. Ejemplo 2 de aplicación del índice

Pregunta

Las benzodiazepinas:

a) son psicofármacos de amplio uso.
b) bajan el umbral convulsivante.
c) tienen efectos hipnóticos.
d) pueden provocar ataxia.
e) provocan relajación muscular.

Índice

	Presenta viñeta	Enunciado Completo	Evita uso de negociaciones	Concordancia gramatical entre enunciado y opciones	Distractores verosímiles	Extensión similar entre las opciones	Evita el uso de Ninguna y Todas las anteriores	Opciones ordenadas	Opciones homogéneas	Aplicación de conocimientos o superior	Puntaje
	0	0	1	1	1	1	1	n/c	0	0	1

n/c = No corresponde

preguntas publicadas del Examen Médico Nacional, el índice resultante fue 3,4.

Todo esto contrasta con una muestra de preguntas de la National Board of Medical Examiners, en los Estados Unidos (para el licenciamiento de médicos), con un índice promedio de 4,5. Esta entidad es considerada por expertos como una de las instituciones que produce exámenes de calidad con preguntas bien construidas. La experiencia adquirida en el uso del índice indica que se debiera intentar al menos obtener un "3" en una pregunta a nivel de asignaturas, pero sobre "4" en exámenes de alta importancia como los usados en licenciamiento. Galli et al.[7], en un análisis de la calidad de cien preguntas de un examen para un programa de actualización pediátrica en Argentina usando el presente índice, encontraron que el 84% de las preguntas tenían una calidad aceptable usando este criterio. Estos autores usan el índice como herramienta para elevar la calidad de las preguntas en este examen, en sus versiones más recientes.

El índice de calidad sirve además como un mecanismo de feedback formativo en talleres y sesiones para mejorar la calidad de la producción de preguntas de opción múltiple. Wallach et al.[3] describen un proceso en una escuela de medicina en

que se midió el índice promedio de una muestra de preguntas con un resultado de 2,5; usando el índice de Josefowicz et al. Luego de establecer pautas claras usando las recomendaciones de Case y Swanson, el índice promedio en años posteriores subió hasta 3,6.

En Chile se ha usado como feedback en algunos talleres de elaboración de preguntas siendo un eficaz mecanismo para comunicar los errores de la pregunta y enfatizar cuáles son las características que necesariamente debe tener una buena pregunta, lo que no impide mencionar, además, las otras características señaladas en la literatura.

Al usar el índice se puede determinar la media de pruebas en un curso o incluso para todo un año. Luego, se puede ir midiendo el progreso alcanzado después de intentar mejorar las preguntas en cada prueba posterior.

Esta medición del índice mediante el formulario puede hacerla cada docente para sus preguntas o el director de curso para ver los potenciales defectos de las preguntas. Con esta información se pueden corregir oportunamente los defectos encontrados.

Entre las limitaciones del índice está el que no indica si la pregunta se relaciona con el objetivo educacional que se desea medir. Este paso es una tarea aparte ya que nuestra intención fue producir un índice que puede ser aplicado en forma independiente por un evaluador, solamente usando el ítem aisladamente en cuanto a su construcción. Sin embargo, en el contexto de los talleres de elaboración de preguntas, se ha incorporado para hacer énfasis en que las preguntas de una prueba tienen que relacionarse con los objetivos del curso.

Estamos conscientes que algunos factores que hemos incorporado en el índice pudieran ser reemplazados por otros. Nuestra intención fue escoger aquellos que fueran representativos de los defectos de construcción mencionados en la literatura, sin tratar de incluirlos todos, para lograr un índice que fuera relativamente simple de calcular.

Creemos que el uso de este índice en las escuelas de ciencias de la salud contribuirá a lograr mejores evaluaciones, donde en la respuesta correcta o incorrecta por parte del estudiante sólo influyan sus conocimientos y no agentes externos como suele suceder cuando las preguntas están mal construidas.

BIBLIOGRAFÍA

1. Josefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D y Glew RH. The quality of In-house medical school examinations. Acad Med 2002; 77(2): 156-161.
2. Case S y Swanson D. Cómo elaborar preguntas para evaluaciones escritas en el área de ciencias básicas y clínicas. Disponible en: www.nbme.org/PDF/IWG-Sp/IWG-Spanish2006.pdf. [Consultado el 10 de septiembre de 2010].
3. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, y Swanson DB. Use of a committee review process to improve the quality of course examinations. Adv Health Sci Educ 2006; 11(1): 61-68.
4. Haladyna TM, Downing SM, Rodriguez MC. A review of item-writing Guidelines for classroom assessment. Appl Meas Educ 2002; 15(3): 309-334.
5. Haladyna TM. Developing and validating multiple-choice test items. Third Edition, Lawrence Erlbaum Associates, New Jersey, 2004.
6. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ 2005; 10(2): 133-143.
7. Galli A, Castrillón S, Martinito R, Maza I, Nakab A y Ageitos ML. Calidad de un examen según escala Galofré. Presentado en el V Congreso de Educación en Ciencias de la Salud, Valdivia, Chile, 2010. Disponible en: www2.udec.cl/ofem/recs/anteriores/vol712010/artval71tredos.htm. [Consultado el 10 de septiembre de 2010].

Correspondencia:
Alberto Galofré T., MD, MEd
Profesor emérito,
Saint Louis University School of Medicine,
640 Broadmoor Dr
St. Louis, MO 63017
St. Louis, Missouri, USA
E-mail: galofrea@slu.edu