Inv Ed Med 2013;**2(8)**:202-210

Investigación en Educación Médica

www.elsevier.es

ARTÍCULO ORIGINAL

Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias?

Alma Jurado-Núñez,[1] Fernando Flores-Hernández,[2] Laura Delgado-Maldonado,[3] Hermann Sommer-Cervantes,[2] Adrián Martínez-González,[2,4] Melchor Sánchez-Mendiola[2,5]

[1] Programa de Apoyo y Fomento a la Investigación Estudiantil (AFINES), Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., México.
[2] Secretaría de Educación Médica, Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., México.
[3] Facultad de Psicología, Universidad Nacional de Educación a Distancia España, México D.F., México.
[4] Departamento de Salud Pública, Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., México.
[5] División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., México.

Recepción 26 de mayo de 2013; aceptación 19 de agosto de 2013

PALABRAS CLAVE
Evaluación del aprendizaje; exámenes escritos; preguntas de opción múltiple; educación médica de pregrado; México.

Resumen
Introducción: Los exámenes de opción múltiple son la herramienta más utilizada en la evaluación del conocimiento en estudiantes de medicina. Se ha demostrado que tres opciones para cada ítem son suficientes, sin embargo, muchos exámenes en nuestro medio aún están compuestos por preguntas con cinco opciones. El estudio de los distractores no funcionales (DNF) es necesario para mejorar la calidad de los exámenes.
Objetivo: Identificar los DNF y su comportamiento en una evaluación sumativa de altas consecuencias en estudiantes de medicina.
Método: Se realizó análisis psicométrico del Examen Profesional Teórico de la Facultad de Medicina de la Universidad Nacional Autónoma de México (UNAM), en la versión de 2008. Se calcularon dificultad, discriminación y correlación punto-biserial de cada ítem y de sus cuatro distractores. Se obtuvo la frecuencia de preguntas con cero a cuatro DNF y se valoraron las diferencias de DNF por ítem y sus características psicométricas. Se contrastó el

Correspondencia: Melchor Sánchez Mendiola. Secretaría de Educación Médica. Av. Universidad N° 3000, Edif. B, 3er piso, C.U., C.P. 04510, México D.F., México. Teléfono: (5255) 5623 2448. Fax: (5255) 5616 2346. Correos electrónicos: melchorsm@gmail.com melchors@liceaga.facmed.unam.mx

comportamiento psicométrico del examen completo, con una versión en la que se eliminaron los ítems con cuatro DNF.
Resultados: El examen tuvo 420 reactivos de opción múltiple con cinco opciones de respuesta, fue contestado por 882 sustentantes. El instrumento tuvo un alfa de Cronbach de 0.93. De los 1 680 distractores evaluados, sólo 788 (46.9%) fueron funcionales. Más de dos tercios del total de ítems contaron con dos o más DNF. Se encontró un promedio de 2.12 ± 0.99 DNF por ítem, la mayoría de los cuales fueron elegidos por menos de 5% de los sustentantes. A mayor cantidad de DNF, mayor índice de dificultad y menor poder de discriminación de los ítems.
Conclusiones: Los ítems con dos distractores funcionales comprenden la mayoría de los reactivos del examen. Dos distractores plausibles representan una alternativa asequible para elaborar los ítems, manteniendo o mejorando la confiabilidad y el perfil psicométrico de este tipo de evaluaciones.

KEYWORDS
Learning assessment; written tests; multiple choice questions; undergraduate medical education; Mexico.

Distractors in multiple-choice questions for medical students: a descriptive analysis in a high-stakes summative exam

Abstract
Introduction: Multiple-choice question (MCQ) exams are the most frequently used tool for knowledge assessment of medical students. It has been shown that three options in each item are enough; however, many tests in our country still use questions with five options. The study of non-functional distractors (NFD) is needed in order to improve the quality of our assessment instruments.
Objective: To identify NFD and their behavior in a high-stakes summative exam in medical students.
Method: A psychometric analysis was performed of the Theoretical Professional Exam at UNAM Faculty of Medicine, the 2008 version. Difficulty, discrimination and point-biserial correlation were calculated for each item and its four distractors. The frequency of items with zero to four NFD was obtained, and the differences of NFD per item and their psychometric characteristics were assessed. The psychometric behavior of the whole exam was compared to a version where items with four NFD were removed.
Results: The test had 420 MCQ items with five options, and was answered by 882 students. The instrument had a Cronbach's alpha of 0.93. Of the 1 680 distractors evaluated, only 788 (46.9%) were functional. More than two-thirds of the total of items had two or more NFD. There was an average of 2.12 ± 0.99 NFD per item, the majority of which were chosen by less than 5% of the test takers. With higher quantity of NFD, the items' difficulty index was higher and the discrimination index was lower.
Conclusions: Items with two functional distractors were the majority in the exam. Two plausible distractors are a reasonable alternative for item development, maintaining or improving the reliability and psychometric profile of these types of tests.

Introducción

Las ciencias de la salud, incluyendo la medicina, consisten en un conjunto complejo de disciplinas que avanzan continuamente en la generación de conocimiento y tecnología. Las instituciones de salud y de educación superior enfrentan los múltiples retos que implican la enseñanza, el aprendizaje y la evaluación de dichos conocimientos y habilidades, en las diferentes poblaciones de estudiantes que son formados en las mismas.

La evaluación del aprendizaje en educación es un proceso continuo, sistemático y reflexivo a través del cual se obtiene información cuantitativa y cualitativa pertinente, válida y fiable acerca de un objeto, lo cual permite identificar fortalezas y áreas de oportunidad para emitir un juicio de su valía o mérito y tomar decisiones fundamentadas orientadas a su perfeccionamiento.[1,2] Existe una gran variedad de instrumentos para evaluar los conocimientos, habilidades y actitudes, todos con sus ventajas y limitaciones. Algunos pueden resultar más útiles o adecuados en determinada situación, porque su eficacia varía según el contexto donde se apliquen, como en diferentes planes de estudios, con grupos o poblaciones de estudiantes de diferentes características, en diversas disciplinas a evaluar, entre otras.[3,4]

Evaluar específicamente el conocimiento adquirido por los estudiantes es una actividad de gran relevancia que permite determinar el logro de los objetivos educativos, así como realimentar a los profesores, a los programas académicos, a la institución y a los propios estudiantes. Para ello, se han utilizado diversos instrumentos de evaluación como preguntas abiertas, exámenes orales, preguntas de opción múltiple (POM), preguntas de verdadero o falso y relación de columnas, entre otros.[5]

En la educación médica de pregrado y posgrado los exámenes que utilizan POM son los instrumentos más frecuentemente utilizados en las evaluaciones de conocimientos.[3,5-7] El uso de este tipo de instrumentos en las pruebas se justifica por su alta aceptación y bajo costo; además, facilita la evaluación de una gran cantidad de estudiantes en poco tiempo, obliga a la estandarización del conocimiento y brinda objetividad y efectividad.[3,5,6] El diseño de las POM permite la evaluación de niveles cognitivos complejos como la aplicación del conocimiento y la solución de problemas. La elaboración de reactivos de forma planeada y siguiendo los principios descritos para elaborar preguntas de calidad, propicia que la evaluación del conocimiento arroje resultados con fuerte evidencia de validez y elevada confiabilidad.[6,8,9]

La visión moderna de validez en evaluación educativa la considera como un concepto holístico, en el que toda la validez es validez de constructo, la cual se nutre de cinco diferentes fuentes de evidencia: contenido, proceso de respuesta, estructura interna, relación con otras variables y consecuencias.[8,10] De acuerdo a este modelo, la fuente de evidencia de validez denominada "estructura interna" de la prueba, se refiere a las características estadísticas o psicométricas del instrumento, de los reactivos que lo conforman, y de las opciones de respuesta.[6,8] En las últimas décadas, diversos investigadores se han dado a la tarea de determinar los elementos clave en la realización de POM y analizar sus características psicométricas, como los índices de dificultad, de discriminación, la correlación punto-biserial, la frecuencia de respuesta de cada una de las opciones, incluyendo a los distractores.[5,6,9,11]

El uso apropiado de distractores u opciones incorrectas es una importante faceta de la elaboración y análisis de los reactivos de opción múltiple, ya que requieren un importante esfuerzo por los elaboradores de ítems, son determinantes de la calidad de las preguntas y tienen una función educativa importante para los sustentantes en el proceso de evaluación.[5,6] La importancia de identificar los distractores y su comportamiento es trascendental para mejorar los estándares de los exámenes, especialmente en aquellos de altas consecuencias; reducir el tiempo para responderlos o ampliar los contenidos a evaluar; igualmente, es relevante para cuestiones administrativas como el tiempo de construcción, y los recursos materiales y humanos implicados en su elaboración. Relativamente pocos trabajos se han publicado sobre el comportamiento de los distractores en los exámenes con POM en ciencias de la salud, que evalúen su calidad, funcionalidad, y análisis psicométrico, lo que hace necesario explorar esta información en instituciones educativas de países como el nuestro.[12-16]

En la Facultad de Medicina de la Universidad Nacional Autónoma de México (UNAM), se forma a los estudiantes durante la licenciatura para que adquieran los conocimientos, habilidades, actitudes y valores necesarios para desempeñarse como médicos generales. Para ello, se cuenta con el Plan de Estudios de la Carrera de **Médico** Cirujano conformado por programas académicos de las áreas biomédica, socio-médica y clínica, estableciendo claramente el perfil del egresado.[17] El Examen Profesional de la Facultad de Medicina de la UNAM es la evaluación sumativa de altas consecuencias aplicada al final de la licenciatura de Médico Cirujano, que tiene como objetivo valorar en conjunto los conocimientos generales del sustentante en su carrera, que éste demuestre su capacidad para aplicar los conocimientos adquiridos y que posee criterio profesional. El examen tiene dos fases, una teórica y otra práctica, la teórica se compone de un examen escrito con reactivos de opción múltiple.[18]

El propósito del presente trabajo es evaluar el comportamiento de los distractores en las POM en una aplicación del Examen Profesional Teórico de la Facultad de Medicina de la UNAM, para obtener información concreta sobre esta temática.

Método

Los datos empleados en el presente estudio corresponden a la información obtenida de los resultados de la Fase Teórica del Examen Profesional de la Facultad de Medicina de la UNAM, aplicado en enero de 2008. En esa ocasión el instrumento se aplicó a la población de estudiantes que acreditaron el quinto año de la carrera de medicina. El instrumento estuvo compuesto por 420 reactivos en formato de opción múltiple, con cinco opciones de respuesta (una correcta y cuatro incorrectas o distractores). La prueba exploró seis áreas de conocimiento: Medicina Interna, Pediatría, Gineco-obstetricia, Urgencias, Cirugía y Medicina familiar. El examen se aplicó en condiciones estandarizadas en instalaciones de la Facultad, las respuestas se registraron en hojas de lector óptico.

La base de datos se analizó mediante la Teoría Clásica de los Test (TCT) a través del programa ITEMAN 3.5 (Assessment Systems Corporation®, Minnesota, EUA www.assess.com), con el que se obtuvieron los datos psicométricos de manera global, de cada ítem y de cada distractor, calculando los índices de dificultad, de discriminación y la correlación punto-biserial, como se reportó previamente.[18]

Los resultados obtenidos se sometieron a un análisis de frecuencia para las opciones de respuesta, identificando qué porcentaje de sustentantes respondió cada una de las opciones. Se definió operacionalmente a los distractores como no funcionales (DNF), tomando en cuenta su desempeño psicométrico, a partir de su asignación a cualquiera de las siguientes categorías: frecuencia menor o igual a 1%, frecuencia menor a 5%, discriminación positiva y no discriminación. Un distractor que es elegido por muy pocos sustentantes es un distractor que no cumple su función evaluativa (puede ser que no sea plausible o que tiene alguna falla en el diseño que hace que los estudiantes lo descarten como opción), y un distractor que no discrimina a los estudiantes o que tiene una discriminación positiva (es decir, es elegido por los mejores estudiantes con mayor frecuencia que por los estudiantes con desempeño más bajo) también es una bandera roja de que ese distractor debe ser revisado por los elaboradores del examen y expertos en contenido.

Se obtuvo el porcentaje de preguntas que contenían de cero a cuatro DNF para cada área de conocimiento y para el examen en general. Se calculó el índice de dificultad y discriminación media para cada grupo de preguntas con número determinado de DNF, para el examen en conjunto y para cada una de las áreas de conocimiento que

integran el examen. Se realizó un ANOVA con una prueba *post-hoc* de Tukey HSD, en que se consideró como factor el número de DNF y como variables dependientes la dificultad, el índice de discriminación y correlación punto-biserial de los ítems; lo anterior con el fin de encontrar diferencias significativas entre grupos y dentro de los mismos. Posteriormente se eliminaron los ítems que presentaron cuatro DNF, y se realizó nuevamente el análisis psicométrico para comparar el desempeño psicométrico de esta versión acortada del examen con el examen completo.

La obtención de datos y análisis de los resultados se llevó a cabo como parte de la evaluación del Plan de Estudios, y sus procesos de control de calidad. Los resultados se manejaron de manera anónima, y se reportan los agregados del análisis, por lo que el trabajo no implica riesgos para los estudiantes evaluados.

Resultados

El Examen Profesional Teórico de la Facultad de Medicina de la UNAM analizado, tuvo lugar en las instalaciones de la institución el mes de enero de 2008, y el número de sustentantes que contestó el examen fue de 882. Respecto al conjunto global de la prueba, se encontró que el promedio de dificultad clásica fue de 54.95% de aciertos, su confiabilidad medida con el coeficiente de Cronbach tuvo un α=0.93 y el error estándar de medición (EEM) de 8.67.

El número total de opciones del examen fue de 2 100, compuesto de 420 respuestas correctas y 1 680 distractores u opciones incorrectas. El 53.1% (892) de los distractores evaluados se clasificó como no funcional (DNF). Solamente 18 ítems (4.3%) tuvieron los cuatro distractores funcionales. Treinta y cuatro ítems (8.1%) no tuvieron ningún distractor funcional. El número de DNF por ítem fue de 2.12 ± 0.99 (media ± desviación estándar). Los resultados de frecuencia se presentan en la **Tabla 1**.

Al clasificar a los DNF, se encontró que 74% de ellos fueron elegidos por menos del 5% de los sustentantes, 36.4% fue respondido por 1% o menos de la población, 20% tuvo un índice de discriminación positiva y 6.1% no discriminó. Cinco áreas del examen contenían, por lo menos, tres ítems con ningún DNF y uno con cuatro DNF. El área de Cirugía, en contraste, tuvo el mayor número de ítems con cuatro DNF y ninguno con todos los distractores funcionales. El comportamiento de DNF en cada área de conocimiento se muestra en la **Tabla 2**.

Tabla 1. Frecuencia de distractores no funcionales (DNF) en el Examen Profesional Teórico de la Facultad de Medicina de la UNAM, aplicación 2008.

DNF por ítem	Total de ítems	Frecuencia (%)
0	18	4.3
1	96	22.9
2	156	37.1
3	116	27.6
4	34	8.1

El promedio de índice de dificultad y de discriminación en los ítems con ningún DNF fue de 0.48 ± 0.14 y 0.30 ± 0.09, respectivamente (media ± desviación estándar). En contraste, estos parámetros psicométricos en los ítems con cuatro DNF fueron 0.84 ± 0.17 y 0.05 ± 0.08, con 40% mayor índice de dificultad (es decir, 40% más fácil) y 25% menor discriminación (**Figura 1**). Para ítems con 2 DNF, los índices de dificultad y discriminación medios fueron 0.52 y 0.20, respectivamente (**Figura 1**).

El ANOVA mostró diferencias estadísticamente significativas (p<0.001) en los indicadores psicométricos (índices de dificultad y discriminación, correlación de punto-biserial), entre los grupos de ítems con diferente número de DNF (**Tabla 3**). La prueba *post-hoc* de Tukey mostró diferencias significativas (p<0.001) en índice de dificultad y de discriminación entre ítems con cuatro DNF y los demás ítems con cero a tres DNF. Se encontró diferencia significativa en la correlación punto-biserial entre los ítems con cuatro DNF y el grupo con ningún DNF (**Tabla 4**).

Al eliminar los 34 ítems con cuatro DNF y analizar de nuevo el examen con la Teoría Clásica de los Test, se mantuvo la confiabilidad y mejoró ligeramente el perfil psicométrico del examen. La **Tabla 5** muestra un pequeño aumento del coeficiente punto-biserial medio, una reducción del error estándar de medición y del índice de dificultad en el examen depurado.

Discusión

El presente trabajo describe una evaluación del comportamiento de todos los distractores utilizados en una aplicación del Examen Profesional Teórico de la Facultad de Medicina de la UNAM, la prueba sumativa de altas consecuencias que se aplica al final de la carrera de médico cirujano en nuestra Institución. Hasta donde pudieron identificar los autores, se trata de uno de los pocos reportes en la literatura disponible, publicada en Latinoamérica, sobre los distractores de POM en exámenes sumativos en escuelas de medicina.

La evaluación en medicina representa un gran reto, especialmente aquella de altas consecuencias; los instrumentos deben ser construidos para evaluar no sólo el conocimiento puntual o memorístico sino niveles cognitivos superiores como la aplicación del conocimiento y la solución de problemas. La redacción de ítems para evaluaciones sumativas con POM es una labor compleja, intensiva, que requiere entrenamiento y práctica y que, en cierto sentido, constituye un arte para los elaboradores de reactivos; las reglas y recomendaciones metodológicas que nutren la elaboración de ítems permanece en continuo desarrollo.[5,6,9,11] El proceso de analizar el desempeño del examen a partir de sus distractores y las características psicométricas de los mismos, es una necesidad imperativa en la mejora continua de la calidad de la evaluación del aprendizaje.

Los distractores deben considerarse una parte importante de los reactivos en los exámenes de opción múltiple, que tienen implicaciones para los evaluadores, los sustentantes y en general todos los usuarios de la información obtenida al aplicar una prueba. Más de medio siglo de investigación en el tema ha identificado una relación sólida entre la elección de los distractores y la puntuación global

Tabla 2. Comportamiento de distractores no funcionales (DNF) por área de conocimiento en el Examen Profesional Teórico de la Facultad de Medicina de la UNAM, aplicación 2008.

Área de conocimiento	Número de ítems con DNF				
	0 DNF	**1 DNF**	**2 DNF**	**3 DNF**	**4 DNF**
Medicina Interna	3	14	27	19	7
Pediatría	4	19	18	22	7
Ginecoobstetricia	3	16	36	13	2
Urgencias	4	14	25	21	6
Cirugía	0	18	22	19	11
Medicina Familiar	4	15	28	22	1
Total	**18 (4.3%)**	**96 (22.9%)**	**156 (37.1%)**	**116 (27.6%)**	**34 (8.1%)**

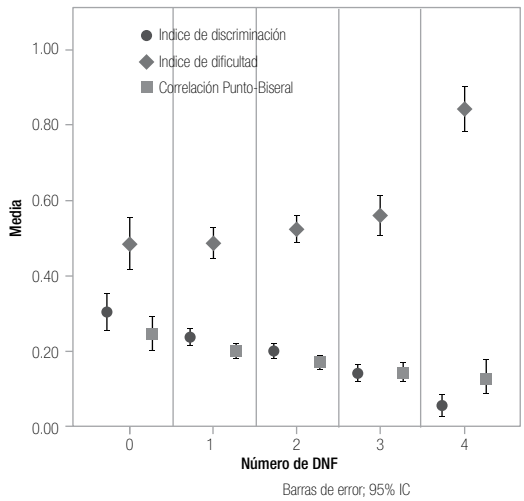

Figura 1. Valores promedio con IC 95% de índice de dificultad, índice de discriminación y correlación de punto-biserial, en cada grupo de ítems con diferente número de distractores no funcionales (DNF), en el Examen Profesional Teórico de la Facultad de Medicina de la UNAM, enero 2008.

en los exámenes, por lo que uno de los expertos mundiales en el área, el Dr. Thomas Haladyna, ha propuesto las siguientes cinco razones para estudiar el desempeño de los distractores en los exámenes de opción múltiple:[19]

- Adelgazar los ítems "obesos".
- Mejorar la calidad de los ítems de los exámenes.
- Detectar las razones que expliquen los problemas de desempeño en los ítems.
- Incrementar los estudios de los procesos cognitivos en el aprendizaje.
- Analizar el funcionamiento diferencial de los distractores.

Estas razones no son sólo convincentes, sino que sugieren que todo el esfuerzo que se aplique en el estudio, análisis y mejora de los distractores, constituirá un componente importante de los argumentos de evidencia de validez para todo el proceso de planeación, diseño, implementación, aplicación y análisis de resultados de los exámenes. En nuestro estudio menos de la mitad de los distractores totales del Examen Profesional Teórico (46.9%) fue funcional, y menos de 5% de los ítems contaron con los cuatro distractores funcionales. Esto sugiere que aunque el examen fue construido con cinco opciones de respuesta (la correcta y cuatro distractores), al aplicar el examen los sustentantes trabajaron generalmente con tres opciones (la correcta y dos distractores). Estos impactantes datos son similares a lo encontrado en otros trabajos, en los que la cantidad de DNF es similar, y la mayoría de los ítems presentan, al menos, un DNF.[12-16,20]

El gran número de ítems con dos DNF coincide con lo mencionado por múltiples autores: tres opciones pueden constituir un probable límite natural y representar la máxima eficiencia en este tipo de reactivos.[12-16,20,21] El uso de cinco o cuatro opciones es aún predominante en evaluaciones en el ámbito médico, a pesar de la gran cantidad de trabajos de investigación publicada sobre el tema y las recomendaciones de los expertos internacionales, por varias razones probablemente relacionadas con los "usos y costumbres" y la tendencia de algunos profesores e instituciones a no conocer o no tomar en cuenta los trabajos de investigación original publicados en las revistas especializadas en educación y medición educativa, en el momento de tomar decisiones sobre cómo enseñar y evaluar.[22] El presente trabajo es un refuerzo más a la necesidad de replantear los estándares que rigen la construcción de nuestros instrumentos, aportando información cuantitativa que puede contribuir a informar las decisiones sobre evaluación que toman los profesores y los gremios institucionales.

Los distractores con frecuencia de selección menor a 5% no son funcionales porque generalmente no son plausibles, resultan ilógicos, carecen de consistencia interna o tienen alguna falla en su redacción que ofrece pistas a los sustentantes.[19,20] La diferencia en el índice de dificultad entre preguntas con 5 y 4 opciones, así como entre 5 y 3 opciones con respecto al ideal (aproximadamente 60%) es de 0.03 y 0.07 respectivamente, por lo que el esfuerzo que implica el desarrollo de distractores extra puede no estar justificado. Las directrices de los principales

Tabla 3. ANOVA de una vía, índices psicométricos por grupos de ítems (cero a cuatro DNF). Examen Profesional Teórico de la Facultad de Medicina de la UNAM, enero 2008.

		Suma de cuadrados	Grados de libertad	Media cuadrática	*F*	*p*
Índice de dificultad	Entre grupos	3.540	4	0.885	16.017	0.000
	Dentro de los grupos	22.931	415	0.055		
	Total	26.472	419			
Índice de discriminación	Entre grupos	1.365	4	0.341	21.619	0.000
	Dentro de los grupos	6.549	415	0.016		
	Total	7.913	419			
Correlación punto-biserial	Entre grupos	0.335	4	0.084	5.080	0.001
	Dentro de los grupos	6.846	415	0.016		
	Total	7.181	419			

expertos internaciones para la construcción de ítems recomiendan la creación de tantos distractores como sea posible y factible, pero además recomiendan establecer tres opciones como objetivo y permitir reactivos con más opciones si se logra su desarrollo con calidad y se confirma su calidad psicométrica.[9,11] En cuanto a la preferencia de los sustentantes, Owen y Froman en 1987 encuestaron a 114 examinados sobre su preferencia entre ítems de tres o cinco opciones; 111 de ellos votaron por tres opciones, tres no externaron preferencia y ninguno votó por cinco opciones.[23]

Los DNF se detectan al ser seleccionados por menos de 5% de los sustentantes, por ser gráficamente similares a la respuesta correcta (es decir, discriminar positivamente) o haber sido contestados igualmente por la población correspondiente a cada percentil de resultados.[20] Algunos estudios que han empleado más de dos distractores, han identificado falta de funcionalidad por lo menos en uno de los excedentes.[12,13,15,20,24] Prácticamente en todos los estudios publicados, incluido el presente trabajo, se ha documentado de manera contundente un porcentaje elevado de DNF en el total de la prueba, lo que es consistente con la hipótesis de que una gran cantidad de distractores en los ítems de exámenes de opción múltiple son desechados con facilidad por los sustentantes, probablemente por ser demasiado obvios, no ser plausibles, o tener pistas que orientan al estudiante a descartarlo, sin necesidad de tener el conocimiento relevante específico supuestamente explorado por el distractor. Es de llamar la atención que todos los estudios publicados que analizan el tema de los distractores, utilizan exámenes sumativos de consecuencias altas o moderadas, en los que se supone se han tomado medidas para cuidar la calidad de los ítems y sus diferentes fuentes de validez durante todo el proceso de elaboración y aplicación de los mismos, y a pesar de ello, el porcentaje de DNF es elevado. Es probable que este fenómeno sea aún más frecuente en exámenes formativos, diagnósticos o de bajas consecuencias, en los que no se tenga tanta atención al detalle durante su integración.

Un argumento utilizado con frecuencia para mantener el exceso de distractores en los ítems es porque así se puede disminuir la probabilidad de respuesta al azar, sin tener que recurrir a fórmulas de corrección.[25-27] Varios autores han demostrado que pocas veces existe más de un distractor funcional por pregunta, lo que puede producir disminución del índice de dificultad y por lo tanto, resultar perjudicial para la discriminación del ítem y para la confiabilidad del examen.[12,13,20] Es importante resaltar que, en la mayoría de los casos, la probabilidad de adivinación se entiende erróneamente o se sobrestima. La probabilidad de adivinar a ciegas o totalmente al azar es la que típicamente se intenta corregir o se considera en el proceso de responder un reactivo del que se desconoce totalmente el contenido; no obstante, se trata de un suceso inverosímil y poco frecuente en sustentantes y estudiantes que, en la vida real, contestan exámenes, especialmente evaluaciones sumativas de altas consecuencias.

El conocimiento parcial, las habilidades para resolver exámenes (*"testwiseness"* en inglés) y la capacidad para discriminar distractores no plausibles (*"educated guessing"* en inglés), son prácticas mucho más frecuentes.[25,28,29] Además, al seguir las directrices para realizar un buen ítem, y con ello lograr un instrumento de evaluación adecuado, se asume que cada pregunta es independiente y cada distractor es funcional, por lo que la probabilidad de acertar ciegamente a una de ellas es 0.33 en ítems de tres opciones, de obtener dos respuestas correctas es 0.11 y así sucesivamente hasta ser extremadamente baja en un examen de más de 100 preguntas, como el evaluado en este trabajo.[29]

La cantidad de preguntas con cuatro DNF es de gran relevancia, ya que pueden amenazar la confiabilidad y validez del examen; especialmente al considerar que se trata de una evaluación de carácter profesional y de altas consecuencias. Estas preguntas tiene una discriminación baja (media = 0.06) y tienen un alto índice de dificultad -es decir, son más fáciles- (media = 0.86), por lo que si se incluyen muchos reactivos con estas características, sustentantes en el límite de aprobación pueden lograr acreditar el examen a expensas de las habilidades referidas para contestar exámenes que no tienen que ver con el conocimiento del constructo explorado (*"testwiseness"*). Como se documentó en este estudio, el eliminar los reactivos con cuatro DNF no afecta la confiabilidad del instrumento, y pudiera mejorar un poco algunos de sus parámetros psicométricos.

Tabla 4. Comparaciones múltiples con el método *post-hoc* de Tukey HSD, de los índices psicométricos por grupos de ítems de cero a cuatro distractores no funcionales (DNF). Examen Profesional Teórico de la Facultad de Medicina de la UNAM, enero 2008.

Variable dependiente	(I) DNF	(J) DNF	Diferencia media (I-J)	Error estándar	Significancia	Intervalo de confianza 95%	
						Límite inferior	Límite superior
Índice de dificultad	Sin DNF	Cuatro DNF	-0.35833	0.06852	0.000	-0.5461	-0.1706
	Un DNF	Cuatro DNF	-0.35927	0.04691	0.000	-0.4878	-0.2307
	Dos DNF	Cuatro DNF	-0.32058	0.04449	0.000	-0.4425	-0.1987
	Tres DNF	Cuatro DNF	-0.28543	0.04584	0.000	-0.4110	-0.1598
	Cuatro DNF	Sin DNF	0.35833	0.06852	0.000	0.1706	0.5461
		Un DNF	0.35927	0.04691	0.000	0.2307	0.4878
		Dos DNF	0.32058	0.04449	0.000	0.1987	0.4425
		Tres DNF	0.28543	0.04584	0.000	0.1598	0.4110
Índice de discriminación	Sin DNF	Dos DNF	0.10464	0.03127	0.008	0.0190	0.1903
		Tres DNF	0.16470	0.03182	0.000	0.0775	0.2519
		Cuatro DNF	0.24944	0.03662	0.000	0.1491	0.3498
	Un DNF	Tres DNF	0.09745	0.01733	0.000	0.0500	0.1449
		Cuatro DNF	0.18219	0.02507	0.000	0.1135	0.2509
	Dos DNF	Sin DNF	-0.10464	0.03127	0.008	-0.1903	-0.0190
		Tres DNF	0.06007	0.01540	0.001	0.0179	0.1023
		Cuatro DNF	0.14481	0.02378	0.000	0.0797	0.2099
	Tres DNF	Sin DNF	-0.16470	0.03182	0.000	-0.2519	-0.0775
		Un DNF	-0.09745	0.01733	0.000	-0.1449	-0.0500
		Dos DNF	-0.06007	0.01540	0.001	-0.1023	-0.0179
		Cuatro DNF	0.08474	0.02450	0.005	0.0176	0.1519
	Cuatro DNF	Sin DNF	-0.24944	0.03662	0.000	-0.3498	-0.1491
		Un DNF	-0.18219	0.02507	0.000	-0.2509	-0.1135
		Dos DNF	-0.14481	0.02378	0.000	-0.2099	-0.0797
		Tres DNF	-0.08474	0.02450	0.005	-0.1519	-0.0176
Correlación punto-biserial	Sin DNF	Tres DNF	0.10440	0.03254	0.012	0.0153	0.1935
		Cuatro DNF	0.11412	0.03744	0.021	0.0115	0.2167
	Un DNF	Tres DNF	0.05877	0.01772	0.009	0.0102	0.1073
	Tres DNF	Sin DNF	-0.10440	0.03254	0.012	-0.1935	-0.0153
		Un DNF	-0.05877	0.01772	0.009	-0.1073	-0.0102
	Cuatro DNF	Sin DNF	-0.11412	0.03744	0.021	-0.2167	-0.0115

Algunas de las limitaciones del presente estudio es que solamente se analizó un Examen Profesional Teórico de una institución, lo que tiene implicaciones para la validez externa y generalizabilidad ecológica de los resultados. Aunque las interpretaciones de los parámetros psicométricos son aseveraciones informadas, existe la posibilidad de que resulten imprecisas por la naturaleza multifactorial de los resultados. El análisis psicométrico del presente trabajo se basó en la Teoría Clásica de los Test, la cual tiene algunas desventajas, como son el hecho de que la dificultad de los ítems está confundida con la habilidad de los estudiantes, su dependencia del tamaño de la muestra y del número de ítems, sus limitaciones para evaluar la aptitud de los sustentantes, entre otras. Una alternativa podría ser utilizar la Teoría de Respuesta al Ítem, ya que ofrece otra perspectiva de análisis psicométrico y considera las limitaciones de la TCT mencionadas.[18] Por otra parte, el estudio sólo evaluó el análisis

Tabla 5. Comparación de análisis de reactivos de la versión completa de la prueba con una versión depurada (sin los 34 ítems con cuatro DNF). Examen Profesional Teórico de la Facultad de Medicina de la UNAM, enero 2008.

Características	Completo	Depurado*
Número de ítems	420	386
Alfa de Cronbach	0.928	0.928
Error estándar de medición	8.67	8.46
p media (dificultad)	0.55	0.52
Coeficiente de punto medio biserial	0.18	0.19
Coeficiente biserial medio	0.25	0.25

* Sin los 34 ítems con cuatro DNF.

psicométrico del instrumento y el análisis estadístico de los distractores que lo conforman, sin evaluar cualitativamente los elementos propios de los mismos (formato, coherencia, longitud, claridad en la redacción, elección de vocabulario), tampoco se realizó una evaluación de los distractores *a posteriori* con expertos en contenido, para identificar las oportunidades educativas de mejora de DNF. Esto será materia de otros trabajos de investigación.

El tópico de la evaluación de los distractores en POM ha recibido relativamente poca atención en el pasado, comparado con otros aspectos de los exámenes. En la actualidad está claro que la búsqueda de evidencia de validez de los instrumentos de evaluación en ciencias de la salud debe tener necesariamente el apartado de "estructura interna", que incluye el análisis psicométrico de los exámenes y el estudio concienzudo del comportamiento numérico y gráfico de todos y cada uno de los distractores que componen los reactivos.[8,19] El uso apropiado de esta información por los elaboradores de exámenes y los usuarios de los resultados puede tener un impacto significativo en la profesionalización del proceso educativo y de evaluación. Algunos de los efectos positivos del análisis de los distractores en este tipo de exámenes son:

- El tener una imagen más realista del instrumento, y no sobreestimar su valor sólo porque tiene muchos reactivos con muchas opciones.
- El disminuir la cantidad de opciones se puede traducir en menor inversión de tiempo en la construcción del examen, tanto de los elaboradores de reactivos como de los responsables del ensamble, aplicación y análisis de la prueba. A los escritores de reactivos les cuesta trabajo generar más de tres o cuatro distractores realmente funcionales, y pueden usar un tiempo excesivo en diseñar una quinta opción que a fin de cuentas tenga defectos y sea fácilmente eliminada por los estudiantes.
- Al disminuir el número de opciones de cada ítem en un examen, el tiempo de resolución del mismo se usa de una manera más eficiente, ya que los sustentantes tardarán menos tiempo en responder un igual número de reactivos.
- Por otra parte, si se cuenta con el mismo tiempo de examinación, se pueden incluir más reactivos. Esto incrementa la validez de contenido al explorar más temas de la tabla de especificaciones (o hacerlo con mayor profundidad).
- Al analizar estadística y gráficamente cada distractor, se encuentra un caudal de oportunidades de mejorar los ítems y en consecuencia, la calidad del instrumento en su conjunto.
- Se identifica el papel educativo de cada distractor, que debe visualizarse desde un punto de vista positivo como una oportunidad de detectar deficiencias en el sistema educativo y en el proceso de enseñanza-aprendizaje, errores de conceptos en los estudiantes, entre otros.
- El distractor es un elemento esencial de los ítems, y por lo tanto debe ser útil para el proceso. Su rol en la exploración de procesos cognitivos de alto nivel es fundamental. Si un distractor no es útil, debe ser eliminado o revisado con minuciosidad.
- Esta información constituye una oportunidad dorada de formación docente para incrementar la conciencia del profesorado sobre lo difícil que es diseñar instrumentos de alta calidad, los elementos para desarrollar distractores funcionales, y la profesionalización gradual de la comunidad de educadores en ciencias de la salud sobre el tema.

El análisis descriptivo reportado en este trabajo arroja información interesante y útil para la comunidad de educadores en ciencias de la salud, que puede contribuir a la mejoría de calidad de los instrumentos e incremento de la validez de las interpretaciones de los resultados. Por otra parte, la información obtenida con este estudio contribuyó a la decisión de utilizar reactivos con menos número de opciones en el examen profesional a partir de 2008, y a seleccionar los reactivos con mejores características psicométricas, manteniendo una confiabilidad adecuada.

Es importante profesionalizar la evaluación del aprendizaje en las escuelas de medicina y demás ciencias de la salud, así como construir reactivos de opción múltiple que sean pertinentes, con distractores educativamente relevantes y sólidos. Cada escuela de medicina, departamento académico y grupo de profesores debe realizar de manera colegiada y lo más profesional posible, el análisis psicométrico de los instrumentos de evaluación que utiliza. La formación de profesores adecuada llevará a mejor redacción de reactivos, mejor diseño de exámenes, mejor medición del aprovechamiento y del nivel de habilidades de los estudiantes, así como a una interpretación más apropiada de los resultados, de los logros y una toma de decisiones más válida para beneficio de los estudiantes, de los educadores y de la sociedad.[4,5]

Contribución de los autores

AJN, FFH, LDM, HSC, AMG y MSM participaron en la concepción y el diseño del estudio, colección de datos, búsqueda bibliográfica y redacción del documento.

AJN, FFH y LDM realizaron el análisis psicométrico de los datos.

Financiamiento

Ninguno.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Presentaciones previas

Trabajo oral en las Jornadas de Educación Médica 2013, Facultad de Medicina de la UNAM.

Referencias

1. Downing SM, Yudkowsky R. Introduction to Assessment in the Health Professions. En: Downing SM, Yudkowsky (Eds). Assessment in Health Professions Education. New York, NY: Routledge; 2009. p. 1-21.
2. Martínez GA, Lifshitz GA, Ponce RR, et al. Evaluación del desempeño docente en cursos de especialización médica. Validación de un cuestionario. Rev Med Inst Mex Seguro Soc 2008;46(4):375-382.
3. Linn RL, Gronlund NE. Measurement and assessment in teaching. 8th ed. USA: Prentice-Hall. 2000. Cap. 2:29-49.
4. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(Suppl):S63-S67.
5. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, Van der Vleuten C, Newble DI (eds). International Handbook of Research in Medical Education. Volume II. Dorcrecht: Kluwer Academic Publishers; 2002. p. 647-672.
6. Haladyna TM. Developing and Validating Multiple-Choice Test Items. 3rd Ed. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Publishers. 2004; Chapter 1:3-18.
7. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach 2004; 26(8):709-712.
8. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ 2003; 37:830-837.
9. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002;15(3):309-334.
10. Messick S. Validity. In: Linn RL (ed). Educational Measurement. 3rd ed. New York: American Council on Education and Macmillan; 1989. p. 13-104.
11. Moreno R, Martínez RJ, Muñiz J. Directrices para la construcción de ítems de elección múltiple. Psicothema 2006;16(3):490-497.
12. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Medical Education 2009; 9:40.
13. Rogausch A, Hofer R, Krebs R. Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: a simulation and survey. BMC Medical Education 2010;10:85.
14. Galli A, Roiter H, de Molleinet D, et al. Evaluación de la calidad de las preguntas de selección múltiple utilizadas en los exámenes de Certificación y Recertificación en Cardiología en el año 2009. Revista Argentina de Cardiología 2011;79(5):419-422.
15. Cizek GJ, O'Day DM. Further investigation of nonfunctioning options in multiple-choice test items. Educ Psychol Meas 1994;54:861-887.
16. Abad FJ, Olea J, Ponsoda V. Analysis of the optimum number alternatives from the Item Response Theory. Psicothema 2001;13(1):152-158.
17. Sánchez-Mendiola M, Durante-Montiel I, Morales-López S, et al. Plan de Estudios 2010 de la Facultad de Medicina de la Universidad Nacional Autónoma de México. Gac Méd Méx 2011;147(2):152-158.
18. Delgado-Maldonado L, Sánchez-Mendiola M. Análisis del examen profesional de la Facultad de Medicina de la UNAM: Una experiencia de evaluación objetiva del aprendizaje con la Teoría de Respuesta al Ítem. Inv Ed Med 2012;1(3):130-139.
19. Haladyna TM. Validity evidence coming from statistical study of item responses. In: Haladyna TM (ed). Developing and Validating Multiple-Choice Test Items. 3rd Ed. Chapter 9. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Publishers; 2004. p. 202-229.
20. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas 1993;53(4):999-1010.
21. Rodríguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract 2005;24(2):3-13.
22. Sánchez-Mendiola M. Educación médica basada en evidencias: ¿ser o no ser? Inv Ed Med 2012; 1(2):82-89.
23. Owen SV, Froman RD. What's wrong with three-option multiple choice items? Educ Psychol Meas 1987;47:513-522.
24. Cizek GJ, Robinson L, O'Day DM. Nonfunctioning options: A closer look. Educ Psychol Meas 1998;58(4):605-611.
25. Rogers WT, Harley D. An empirical comparison of three- and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. Educ Psychol Meas 1999;59:234-247.
26. Bruno JE, Dirkzwager A. Determining the optimal number of alternatives to a multiple-choice test. Educ Psychol Meas 1995;55(6):959-966.
27. Burton RF. Multiple choice and true/false tests: reliability measures and some implications of negative marking. Assessment & Evaluation in Higher Education 2004;29(5):585-595.
28. Burton RF. Multiple-choice and true/false tests: myths and misapprehensions. Assessment & Evaluation in Higher Education 2005;30(1):65-72.
29. Downing SM. Guessing on selected-response examinations. Med Educ 2003;37:670-671.