Iatreia
ISSN: 0121-0793
revistaiatreia@udea.edu.co
Universidad de Antioquia
Colombia

Salazar Blanco, Olga Francisca; Marcela Vélez, Claudia; Zuleta Tobón, John Jairo
Evaluación de conocimientos con exámenes de selección múltiple: ¿tres o cuatro opciones de respuesta? Experiencia con el examen de admisión a posgrados médico-quirúrgicos en la Universidad de Antioquia
Iatreia, vol. 28, núm. 3, julio-septiembre, 2015, pp. 300-311
Universidad de Antioquia
Medellín, Colombia

Evaluación de conocimientos con exámenes de selección múltiple: ¿tres o cuatro opciones de respuesta? Experiencia con el examen de admisión a posgrados médico-quirúrgicos en la Universidad de Antioquia

Olga Francisca Salazar Blanco[1], Claudia Marcela Vélez[2], John Jairo Zuleta Tobón[3]

RESUMEN

Introducción: el objetivo de este estudio fue evaluar el efecto de la reducción del número de opciones de respuesta por pregunta sobre los indicadores sicométricos de un examen de ingreso a estudios médicos de posgrado.

Metodología: aplicación de índices de evaluación sicométrica desde la perspectiva de dos teorías: la clásica de la medición y la de respuesta al ítem, a una prueba de 70 preguntas hecha a 2.539 aspirantes a ingresar a los posgrados médico-quirúrgicos de la Universidad de Antioquia en el año 2014. Se eliminó la opción de respuesta elegida con menor frecuencia y se la reemplazó por azar de entre las tres restantes.

Resultados: solo 52,9% de las preguntas tuvieron tres opciones funcionales de respuesta. No se encontró diferencia en la dificultad, la discriminación, el error estándar de la medición, el alfa de Cronbach ni el coeficiente de correlación biserial (teoría clásica de la medición); tampoco en la medida de dificultad de los ítems o de habilidad de las personas (teoría de respuesta al ítem) entre las pruebas con tres y cuatro opciones de respuesta. La prueba con tres opciones conservó un buen ajuste.

Conclusión: una prueba con tres opciones de respuesta se comportó tan bien como su contraparte de cuatro opciones.

[1] Profesora, Departamento de Pediatría y Puericultura. Coordinadora Académica, Grupo de Investigación EDUSALUD, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia.

[2] Médica, especialista en Salud Pública, estudiante de Maestría en Ciencias Clínicas, Grupo Académico de Epidemiología Clínica (GRAEPIC). Profesora de la Facultad de Medicina, de la Universidad de Antioquia, Medellín, Colombia.

[3] Profesor de Ginecología y Obstetricia, Magíster en Epidemiología Clínica. Universidad de Antioquia, Medellín, Colombia.
Correspondencia: John Jairo Zuleta Tobón; jjzuleta@une.net.co

Recibido: febrero 5 de 2015
Aceptado: marzo 20 de 2015

Cómo citar: Salazar Blanco OF, Vélez CM, Zuleta Tobón JJ. Evaluación de conocimientos con exámenes de selección múltiple: ¿tres o cuatro opciones de respuesta? Experiencia con el examen de admisión a posgrados médico-quirúrgicos en la Universidad de Antioquia. Iatreia. 2015 Jul-Sep;28(3): 300-311. DOI 10.17533/udea.iatreia.v28n3a08.

PALABRAS CLAVE

Análisis de Ítem; Evaluación Educacional; Pruebas de Selección Múltiple

SUMMARY

Evaluation of knowledge with multiple-choice tests: three of four options? Experience with admission examinations to medical and surgical postgraduate studies at University of Antioquia (Medellín, Colombia)

Introduction: The aim of this study was to evaluate the effect of reducing the number of response options per question on the psychometric indicators of an exam for admission to postgraduate medical studies, at University of Antioquia, in Medellín, Colombia.

Methodology: Application of psychometric assessment indexes from the perspective or two theories: the classical of measurement and the item response, to a test of 70 questions, applied in 2014 to 2.539 candidates. The least frequently chosen distractor was eliminated and randomly replaced by one of the three remaining ones.

Results: Only 52.9% of the questions had three functional distractors. No difference was found in the difficulty, discrimination, standard error of measurement, Cronbach's alpha and the coefficient of biserial correlation (classical measurement theory). Also, there was no difference in the extent of item difficulty or ability of people (item response theory). The test with three options retained a good fit.

Conclusion: Multiple choice tests with three response options performed as well as their four options counterparts.

KEY WORDS

Educational Measurement; Item Analysis; Multiple Choice Questions

RESUMO

Avaliação de conhecimentos com exames de seleção múltipla: três ou quatro opções de resposta? Experiência com o exame de admissão a pós-graduações médico-cirúrgicos na Universidade de Antioquia

Introdução: o objetivo deste estudo foi avaliar o efeito da redução do número de opções de resposta por pergunta sobre os indicadores psicométricos de um exame de rendimento a estudos médicos de pós-graduação.

Metodologia: aplicação de índices de avaliação psicométrica desde a perspectiva de duas teorias: a clássica da medição e a de resposta ao item, a uma prova de 70 perguntas feita a 2.539 aspirantes a ingressar às pós-graduações médico-cirúrgicos da Universidade de Antioquia no ano de 2014. Eliminou-se a opção de resposta eleita com menor frequência e se a substituiu por casualidade de entre as três restantes.

Resultados: só 52,9% das perguntas tiveram três opções funcionais de resposta. Não se encontrou diferença na dificuldade, a discriminação, o erro regular da medição, o alfa de Cronbach nem o coeficiente de correlação bisserial (teoria clássica da medição); também não na medida de dificuldade dos itens ou de habilidade das pessoas (teoria de resposta ao item) entre as provas com três e quatro opções de resposta. A prova com três opções conservou um bom ajuste.

Conclusão: uma prova com três opções de resposta se comportou tão bem como sua contraparte de quatro opções.

PALAVRAS CHAVE

Análises de Item; Avaliação Educacional; Provas de Seleção Múltipla

INTRODUCCIÓN

La evaluación, en su acepción pedagógica, es amplia porque está en relación con cualquier proceso por medio del cual se analizan una o varias características de un estudiante, de un grupo de ellos o de un ambiente educativo para valorarlos de acuerdo con unos criterios o puntos de referencia con el fin de emitir un juicio (1). La selección es uno de los fundamentos de la evaluación educativa, presente desde las prácticas chinas de selección extraescolar y posteriormente en la Edad Media, para evitar las presiones burocráticas

(1). En la actualidad es común que se hagan pruebas de selección de aspirantes a programas de pregrado y posgrado en las instituciones de educación superior, algunas orientadas exclusivamente a los conocimientos, y otras, que pretenden evaluar más integralmente al aspirante, incluyen también las habilidades y actitudes, de acuerdo con el programa. Las pruebas de evaluación de selección múltiple son los instrumentos más estudiados y más usados, y se ha demostrado que ofrecen buen poder discriminativo, buena confiabilidad o fiabilidad, características de la validez de una prueba de gran importancia para señalarla como objetiva, con la ventaja logística y de costos para aplicarla a un número grande de evaluados (2-4).

Uno de los elementos más discutidos en la elaboración de las pruebas de selección múltiple es el número de opciones de respuesta adecuado para medir de manera confiable el conocimiento sin que se alteren la complejidad ni el poder discriminativo (5). Es frecuente que en las pruebas con cuatro o cinco opciones de respuesta se incluyan una o dos alternativas muy obvias o poco razonables por el simple hecho de cumplir con la directriz general del número de opciones (6). Estas opciones aumentan el tiempo de elaboración de la prueba por parte del docente y el de lectura para el evaluado. Las pruebas con tres opciones de respuesta ofrecen la ventaja de ser más fáciles de construir para los docentes, con menor riesgo de incluir alternativas inadecuadas y menor tiempo de lectura, lo cual posibilita el aumento del número de preguntas con lo que se logra una mayor cobertura de contenidos y mayor confiabilidad o reproducibilidad de la prueba (7). Más por tradición que por demostraciones objetivas, algunos siguen recomendando el uso de cuatro opciones, con el argumento de que tienen mayor complejidad y mejor poder discriminativo y con este concepto se construyen las pruebas de selección múltiple de los exámenes del Estado, como ICFES y ECAES, y las de las universidades, incluido el examen de admisión a los posgrados médico-quirúrgicos de la Facultad de Medicina de la Universidad de Antioquia.

El número de opciones de respuesta de un examen de selección múltiple puede variar dependiendo del escenario educativo y del tipo de evaluación que se pretenda hacer. Algunos autores han demostrado que las preguntas con tres opciones tienen un adecuado poder de discriminación (6), sin embargo, la evaluación de la pertinencia de tres, cuatro o cinco opciones de respuesta para las preguntas de selección múltiple se ha hecho predominantemente en exámenes no médicos (8). En Colombia, según el conocimiento de los autores, no existen publicaciones sobre este tema en áreas de la salud; por lo tanto, el objetivo de este estudio fue evaluar los cambios en las características de la prueba, los resultados y las decisiones al pasar de un instrumento de selección múltiple con cuatro a uno con tres opciones de respuesta en un proceso de evaluación de médicos generales. Esta sería una buena opción para próximos exámenes de admisión, generalizable a otras pruebas en la Facultad y a otras facultades y universidades del área de la salud del país.

METODOLOGÍA

Estudio descriptivo para el cual se utilizaron dos enfoques teóricos de la medición: la teoría clásica y la teoría de respuesta al ítem. Se tuvieron en cuenta los exámenes de 2.539 aspirantes a ingresar a 21 programas de posgrado clínico-quirúrgicos de la Facultad de Medicina de la Universidad de Antioquia, Medellín, Colombia, en el año 2014. Los datos se utilizaron de manera anónima, identificados por un código y la única información que se tomó fue la universidad donde cada candidato terminó el pregrado, y el posgrado al cual aspiraba; por lo tanto, esta investigación no requirió evaluación de un Comité de Ética.

La prueba consta de 70 preguntas de selección múltiple constituidas por un tallo con la descripción de un caso clínico de cualquiera de las especialidades a las cuales aspiran los evaluados y cuatro alternativas de respuesta, que pueden ser de dos tipos: el primero con una sola respuesta verdadera, y el segundo, con todas las opciones de respuesta verdaderas, pero con una de ellas más adecuada que el resto para la situación clínica específica (9).

Las preguntas las elaboran los profesores de las diferentes especialidades siguiendo unas directrices: incluir temas pertinentes para el perfil epidemiológico de Colombia y para las condiciones clínicas que con

mayor frecuencia encuentran los médicos generales que ejercen en cualquiera de los servicios institucionales: urgencias, consulta externa, hospitalización de bajo nivel de complejidad, programas de promoción y prevención o atención primaria. Una comisión de cuatro profesores con experiencia en elaboración de pruebas evalúa una a una las preguntas con el fin de garantizar su validez. Se intenta que estos casos clínicos exijan la integración de diferentes dominios del aprendizaje, se requiere la memoria o recuerdo, pero también la comprensión de los conceptos, su aplicación y análisis, es decir, cuatro de los seis objetivos del dominio cognoscitivo del aprendizaje humano de la teoría de Bloom susceptibles de evaluación con este tipo de preguntas (10,11). En este proceso, se dejan en el tallo de la pregunta los elementos estrictamente necesarios para entender la situación problemática que se presenta; en caso necesario se mejora la redacción para que quede bien desde los puntos de vista gramatical y ortográfico, además de clara y concisa; se privilegian las preguntas positivas y se dejan como negativas solo aquellas en las que realmente tiene utilidad clínica que el aspirante conozca un aspecto negativo; se evita que haya trucos o aspectos diferenciadores para la respuesta correcta que no sean los verdaderamente importantes desde el punto de vista clínico. Con respecto a las opciones de respuesta, se busca que sean alternativas incorrectas, pero que parezcan admisibles, es decir, que no sean descartadas de manera obvia, sino que tengan la posibilidad de atraer a los aspirantes que tienen menos, pero no a los que tienen más conocimiento del tema; que sean de igual extensión y con forma y estilo gramatical similares, concordantes con la pregunta, que no den claves de respuesta para esa o para otra pregunta; ordenadas de manera aleatoria o en un orden lógico cuando la pregunta lo amerite (numérico, por pasos); sin doble negación y sin alternativas como *ninguna* o *todas las anteriores* o con combinación de opciones (ejemplo: "a y c") (2). Como opciones de respuestas incorrectas se privilegian alternativas de las que se sabe que existe evidencia en contra, pero no es infrecuente que en la práctica diaria las aplique un médico con conductas inadecuadas o desactualizadas, es decir, se sigue la recomendación de expertos para la elaboración de las opciones de respuesta: "para redactar los *distractores* use los errores típicos de los estudiantes"(2), cada *distractor* debe utilizar las ideas erróneas comunes con respecto a la respuesta correcta (3,4,6).

Análisis según la teoría clásica de medición

Se calculó la dificultad de cada pregunta mediante la proporción de aspirantes que la respondieron de manera correcta, según la respuesta asignada por el especialista que la elaboró y corroborada por el comité de preguntas. Algunos autores cuestionan el nombre de "dificultad" y proponen en su reemplazo el de "facilidad", porque a mayor valor del índice de dificultad, más fácil es la pregunta. Se calculó la proporción de aspirantes que optó por cada una de las opciones incorrectas y se tomó el umbral que con más frecuencia se propone en la literatura para considerar una opción de respuesta como no funcional: que sea elegida por menos de cinco por ciento de los evaluados (8,12). Se evaluó la discriminación mediante el índice y el coeficiente de discriminación. El índice de discriminación de cada pregunta se calculó como el número de respuestas correctas dadas por el 27% de aspirantes con mejor resultado global de la prueba menos el número de respuestas correctas dadas por el 27% de aspirantes con los resultados inferiores, dividido por el mayor número de aspirantes en uno de estos grupos, en este caso 686 del grupo de rendimiento inferior (en el 27% superior quedaron 685 aspirantes) (13). Este índice fluctúa entre -1 y +1 y cuanto mayor sea su valor, mayor es su capacidad para diferenciar entre los evaluados con calificaciones altas y bajas. Se categorizó según la recomendación frecuentemente asignada en la literatura a Ebel: menor de cero, pésima; de cero a 0,2, pobre; de 0,2 a 0,29, regular; de 0,3 a 0,39, buena; y mayor de 0,39 excelente (14,15). Para el cociente de discriminación se utilizó el coeficiente de correlación de punto biserial que representa la correlación entre cada pregunta y el resultado total de la prueba, que además de tener en cuenta los resultados de todos los evaluados excluye la pregunta que se está evaluando del puntaje global del evaluado. Este coeficiente evalúa qué tanto una pregunta predice el resultado global de la prueba o, lo que es lo mismo, si los mejor evaluados son los que contestan de manera correcta las preguntas (13). Se calculó la confiabilidad del examen mediante el alfa de Cronbach, que es una medida de la reproducibilidad, en este caso utilizada para evaluar la consistencia interna de la prueba,

cuyo valor fluctúa entre 0 y 1. Un valor alto significa que si se repitiera la misma prueba a los mismos evaluados sin que cambiaran las condiciones (entre ellas que los evaluados no aprendieran u olvidaran nada en el período entre las evaluaciones), obtendrían resultados similares (16). Una interpretación alternativa es la probabilidad de que dos personas con el mismo nivel de habilidad o conocimiento obtengan la misma calificación en la prueba.

En evaluaciones para la toma de la decisión de pasar o no un umbral, la consecuencia de resultados falsos positivos o falsos negativos es más importante que el valor absoluto del coeficiente de confiabilidad; por lo tanto, se ha propuesto el índice de reproducibilidad paso/falla (*Pass/failure reproducibility index*), que estima el grado de confianza que se tiene en la decisión tomada con los resultados de la evaluación (16). Este índice fluctúa entre 0 y 1 y evalúa la probabilidad de tomar la misma decisión si se repitiera la prueba. Cada programa tiene su propio punto de corte, pero para hacer este análisis se tomaron los primeros 108 puntajes (número total de cupos en esta convocatoria) para evaluar la concordancia de la decisión entre la prueba con tres o con cuatro opciones de respuesta y para ello se utilizaron las tablas de Subkoviak (17).

Teoría de respuesta al ítem

El análisis de las pruebas con la teoría de respuesta al ítem, y específicamente con la metodología Rasch, tiene como ventaja sobre el análisis con la teoría clásica de medición que tiene en cuenta de manera simultánea el nivel de conocimiento de quienes toman la prueba y el nivel de dificultad de las preguntas, con lo cual se logra que los resultados sean independientes de la habilidad de la población estudiada, situación que no sucede con los resultados de los análisis de teoría clásica, en los que no se puede diferenciar ni separar la habilidad de las personas de la dificultad de los ítems. Los índices y estadísticos utilizados tanto para personas como para ítems fueron: raíz cuadrada del error medio (MNSQ); estadísticos de ajuste próximo (*infit*) y lejano (*outfit*); separación y confiabilidad; índice de dificultad de los ítems y de habilidad de las personas y confiablidad general de la prueba.

Los índices de dificultad de los ítems y de la habilidad de las personas para contestar correctamente la pregunta se dan en una medida logarítmica (*measure,* por lo general entre -3 y +3). En el caso de los ítems, a mayor negatividad de la medida más fácil es el ítem, y a mayor positividad, mayor es la dificultad. En el caso de la medición de las personas, a mayor negatividad de la medida menor es la habilidad del aspirante para contestar las preguntas, y a mayor positividad mayor es la habilidad del aspirante para contestar correctamente. Resultados cercanos a 0, en ambos casos, significan que el ítem tiene una dificultad media y que el aspirante tiene una habilidad promedio. El mapa de Wright permite identificar gráficamente esta relación entre ítems y personas.

La medida de ajuste lejano (*outfit*) de la persona evalúa el comportamiento inesperado del evaluado en las preguntas alejadas de su nivel de conocimiento, mientras que la medida de ajuste lejano (*outfit*) del ítem mide el comportamiento inesperado de la pregunta en los evaluados alejados del nivel de dificultad de esa pregunta. Esto quiere decir que el modelo Rasch detecta cuántos aspirantes con baja habilidad responden preguntas de alta dificultad y viceversa, y cuántos ítems de baja dificultad son fallados por aspirantes con alta habilidad y viceversa, situaciones que reflejan adivinanza o descuido al contestar. La diferencia entre el *outfit* y el *infit* radica en que el primero evalúa valores extremos, mientras que el segundo tiene su foco centrado en respuestas inesperadas alrededor del promedio. Unos *infit* aceptables indican que las preguntas se ajustan bien al grupo de evaluados para quienes se dirigieron las preguntas y unos *outfit* aceptables indican que la prueba está libre de preguntas redundantes, irrelevantes o dependientes entre ellas. El análisis del ajuste próximo y lejano se hace con base en los residuales del modelo y se presenta con la media cuadrática (*Mean Square* MNSQ) y los índices estandarizados (ZSTD). MNSQ evalúa la precisión de la estimación tanto para personas como para ítems y refleja el "ruido no modelado" u otras fuentes de variabilidad en los datos, para lo cual se acepta que valores entre 0,7 y 1,3 reflejan un buen ajuste. ZSTD muestra la significación estadística de las respuestas inesperadas observadas, y se espera que sus valores estén entre -2 y 2.

La separación de personas e ítems permite evaluar el poder de la medición para discriminar entre aspirantes con diferentes niveles de habilidad, y entre preguntas con diferentes niveles de dificultad. Esto

quiere decir que si la escala separa adecuadamente, es posible diferenciar los aspirantes con baja habilidad, habilidad promedio y alta habilidad, de tal manera que la prueba permita seleccionar efectivamente a los de mayor habilidad, y separar adecuadamente entre ítems de baja dificultad, dificultad promedio y alta dificultad, de tal modo que se excluyan de la prueba las preguntas extremadamente fáciles y extremadamente difíciles. Para ser adecuada, la separación de ítems y personas debe ser de al menos 3 errores estándar, y esta medida se correlaciona con la confiabilidad, medida con el alfa de Cronbach, la cual debe ser mayor de 0,7.

Para identificar diferencias entre una prueba con cuatro y una con tres opciones de respuesta, de cada pregunta se eliminó la opción menos elegida por los evaluados y de manera aleatoria se asignó una nueva respuesta de las tres restantes. Diferentes autores han utilizado esta estrategia (5), la cual asume que el evaluado que no conoce la respuesta correcta elige al azar entre las diferentes opciones (8). Se presentan las medidas de análisis de ítem para cada versión de la prueba y se hizo comparación de la dificultad y la discriminación mediante los límites de acuerdo de Bland-Altman, método gráfico en el que el cero en el eje Y representa un acuerdo perfecto y en el cual para considerar que las mediciones con los dos métodos son bastante similares, y posiblemente intercambiables, más del 95% de las diferencias entre las mediciones por los dos métodos en estudio se deben ubicar entre más y menos dos desviaciones estándar alrededor de la media de la diferencia (límites de acuerdo) (18).

Para los análisis estadísticos se utilizaron los programas Excel, SPSS 21.0, Winsteps 3.70.0 y Epidat 4.0.

RESULTADOS

El examen constó de 70 preguntas, cada una con cuatro opciones de respuesta y fue respondido por 2.539 aspirantes. En 33 preguntas (47,1%) las tres opciones incorrectas fueron funcionales, 29 (41,4%) tuvieron dos opciones incorrectas funcionales, 7 (10%) tuvieron solo una opción incorrecta funcional y una pregunta (1,4%) no tuvo opciones incorrectas funcionales, es decir, el 52,9% de las preguntas tuvieron al menos una opción incorrecta que fue atractiva para menos del 5% de los aspirantes. En la tabla 1 se presenta el análisis de la prueba con cuatro y con tres opciones de respuesta desde la teoría clásica de la medición; la diferencia más importante es el aumento de preguntas con opciones de respuesta funcionales con la prueba de tres opciones, mientras que los diferentes índices y demás parámetros de evaluación son bastante similares entre ellas.

Los índices más importantes para evaluar las pruebas desde esta teoría son la dificultad y la discriminación. Las figuras 1 y 2 presentan la evaluación del acuerdo entre las dos modalidades de la prueba mediante los gráficos de Bland Altman; en ambos casos se observa que menos del 5% de las preguntas están por fuera de los límites de acuerdo y que la diferencia para estos dos índices entre las dos pruebas es numéricamente pequeña (eje Y).

Cada programa recibe un número máximo preestablecido de aspirantes, elegido en orden descendente de calificación entre los aspirantes a él, independientemente de la calificación obtenida con respecto al total de aspirantes. La tabla 2 muestra que si el examen hubiera tenido preguntas con tres opciones de respuesta, en vez de cuatro, la coincidencia en la decisión hubiera sido del 95,4% y solo en 5 casos de los 2.539 (0,2%) se habría tomado una decisión diferente.

La tabla 3 presenta el análisis de las dos modalidades de la prueba desde la teoría de respuesta al ítem con el modelo de Rasch. Aunque no es el objetivo del estudio, se observa que hubo un ajuste adecuado de los datos obtenidos con los modelos, sustentado con los rangos de valores de los índices de ajuste próximo y lejano y, lo importante para el objetivo del estudio, se observa bastante similitud entre las medidas evaluadas.

El mapa de Wright representa la habilidad de las personas y la dificultad de los ítems; se observa que los ítems 13 y 20 fueron los más sencillos y los ítems 47 y 39, los más difíciles (figura 3).

Tabla 1. Descripción de las características de la prueba con tres o con cuatro opciones de respuesta según la teoría clásica de la medición

	Prueba con cuatro opciones	Prueba con tres opciones
Índice de dificultad	51,3 (DS 17,8)	53,2 (DS 17,2)
Índice de discriminación	0,22 (DS 0,12)	0,21 (DS 0,12)
Pobre discriminación (< 0,2)	31 (43,4%)	34 (48,6%)
Discriminación regular (0,2 a 0,29)	21 (30,0%)	20 (28,6%)
Buena discriminación (0,3 a 0,39)	11 (15,7%)	11 (15,7%)
Excelente discriminación (> 0,39)	7 (10,0%)	5 (7,1%)
Coeficiente de correlación biserial de punto	0,12 (0,09)	0,11 (0,09)
Alfa de Cronbach	0,626	0,585
Error estándar de la medición	3,91	3,85
Preguntas con al menos una opción de respuesta no funcional	37 (52,9%)	9 (12,9%)
Preguntas con una opción de respuesta no funcional	29 (41,4%)	8 (11,4%)
Preguntas con dos opciones de respuesta no funcionales	7 (10,0%)	1 (1,4%)
Preguntas con tres opciones de respuesta no funcionales	1 (1,4%)	No aplica
Índice de reproducibilidad paso/falla	0,72	0,73

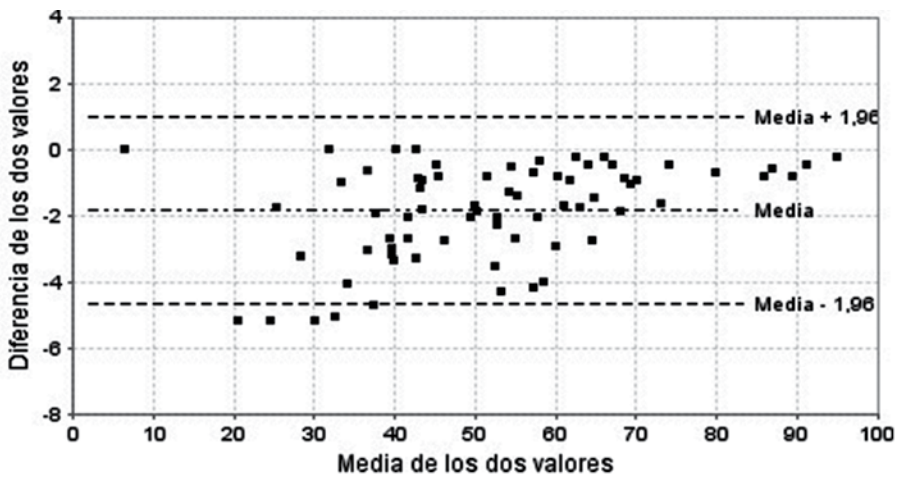

Figura 1. Límites de acuerdo para la dificultad de las pruebas con tres o cuatro opciones de respuesta

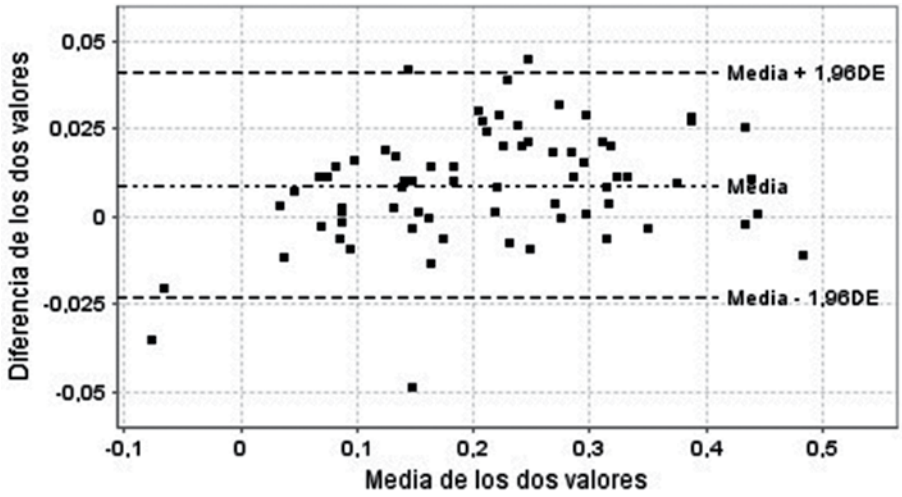

Figura 2. Límites de acuerdo para la discriminación de las pruebas con tres o cuatro opciones de respuesta

Tabla 2. Elección de candidatos con la prueba con tres o cuatro opciones de respuesta

		Cuatro opciones de respuesta		
		Elegido	**No elegido**	**Total**
Tres opciones de respuesta	Elegido	103	5	108
	No elegido	5	2.416	2.421
	Total	108	2.421	2.529

Tabla 3. Descripción de las características de la prueba con tres o con cuatro opciones de respuesta, según la teoría de respuesta al ítem, modelo de Rasch

Medida	**Personas**		**Ítems**	
	Prueba con cuatro opciones	**Prueba con tres opciones**	**Prueba con cuatro opciones**	**Prueba con tres opciones**
Raíz cuadrada del error medio (RMSE)	0,27	0,26	0,05	0,05
Habilidad de las personas/Dificultad de las preguntas. Media (DE)	0,08 (0,42)	0,18 (0,4)	0 (0,9)	0 (0,85)
Rango de dificultad	-1,49 a 2,04	-1,28 a 2,01	-2,93 a 2,86	-2,23 a 1,39
Ajuste *Infit* MNSQ (rango)	0,7;1,31	0,76;1,27	0,92-1,11	0,93-1,09
Ajuste *Outfit* MNSQ (rango)	0,7;1,76	0,72;1,66	0,8-1,15	0,91-1,14
Número de personas/ítems que desajustan	330	297	0	0
Separación	1,25	1,15	19,87	18,77
Confiabilidad (Alfa de Chronbach)	0,61	0,57	1	1

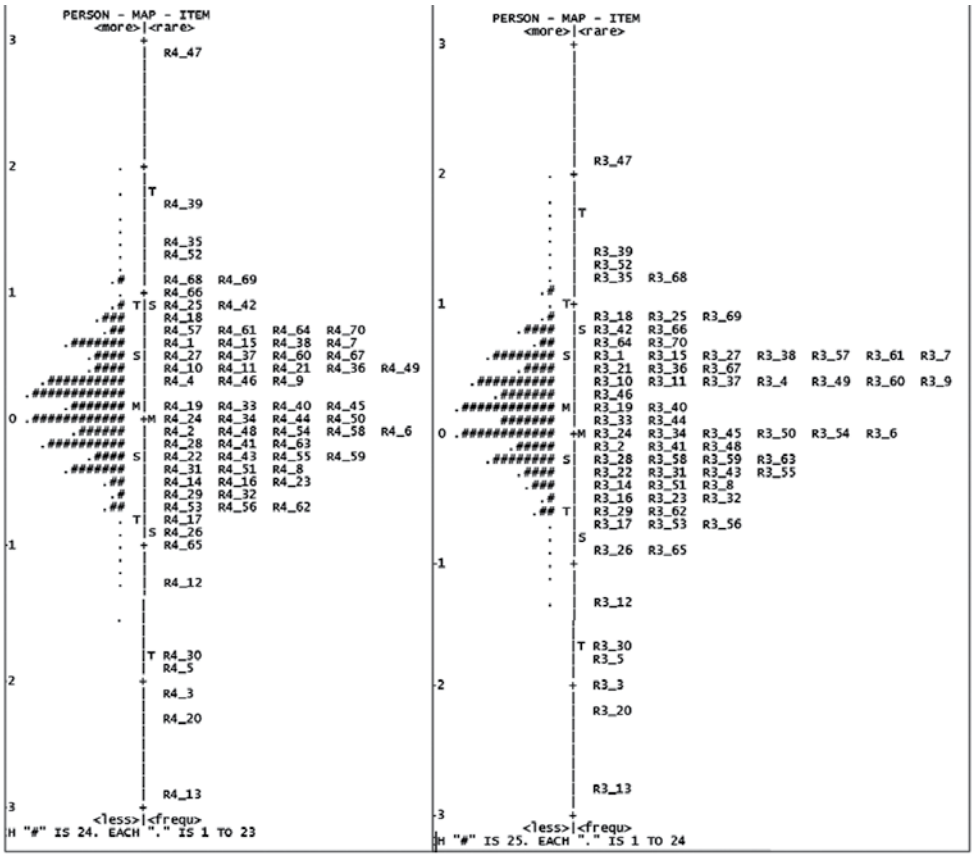

Figura 3. Mapas de Wright para la prueba con cuatro o tres opciones de respuesta

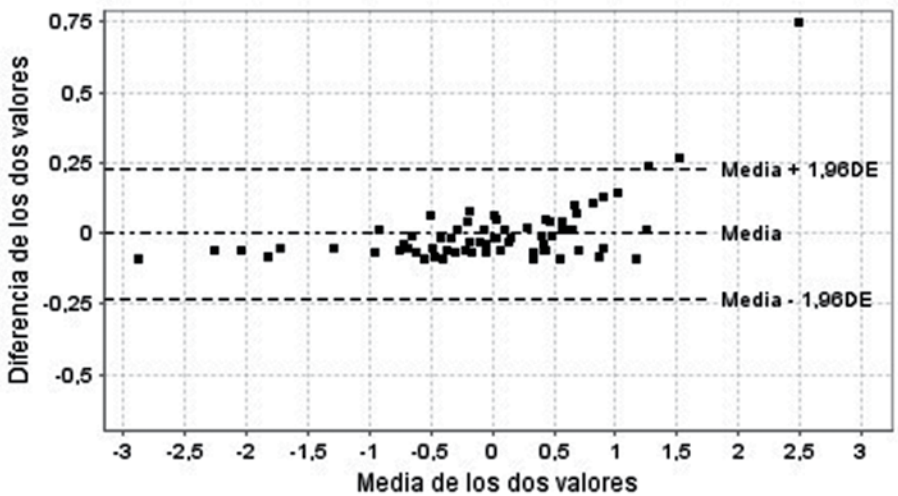

Figura 4. Límites de acuerdo para la dificultad de los ítems con tres o cuatro opciones de respuesta

El número de personas que desajustan en el modelo es diferente, siendo mejor para la prueba con tres opciones. La separación es mejor con la prueba de cuatro opciones, pero la diferencia es muy pequeña. Según esto, tener una opción de respuesta más no le agrega valor a la prueba. La medición de la dificultad de las preguntas es bastante similar, como lo muestra el gráfico de Bland Altman (figura 4).

DISCUSIÓN

El presente estudio, realizado con una muestra grande y representativa de médicos de diferentes universidades y sitios del país, confirma lo encontrado por otros autores en estudiantes o graduados de otras profesiones: una prueba con tres opciones de respuesta no es inferior en dificultad y capacidad de discriminar a los evaluados a una prueba con cuatro opciones.

La propuesta de disminuir el número de opciones de respuesta en las pruebas de evaluación de conocimientos parte del hecho reconocido que con mucha frecuencia una o más de esas opciones no aporta a la dificultad o al poder de discriminación de la prueba. Un estudio que evaluó el funcionamiento de las opciones de respuesta en siete exámenes para estudiantes de enfermería encontró que solo la mitad de las opciones incorrectas de respuesta fueron funcionales y solo en el 13,8% de las 514 preguntas todas las opciones de respuesta incorrectas fueron clasificadas como funcionales (6). En este estudio evaluaron el efecto de disminuir el número de opciones de respuesta incorrecta de tres (en las preguntas con cuatro opciones) a dos (en las que solo tienen tres opciones), y encontraron poca diferencia en la dificultad de las preguntas. Estos hallazgos se reproducen en nuestro estudio, en el que la mitad de las preguntas tuvieron al menos una opción incorrecta que fue elegida por menos del 5% de los aspirantes y no se encontraron diferencias entre los distintos índices y resultados de la prueba con tres o cuatro opciones. La inclusión en un examen de opciones de respuesta no funcionales no mejora la discriminación o el poder de evaluación del conocimiento y en cambio sí aumenta el tiempo de lectura o de análisis y en ocasiones dan claves a los evaluados, con lo que se incrementa la posibilidad de responder correctamente una pregunta sin tener conocimiento del tema (19).

Un metaanálisis con 27 estudios que incluyó pruebas de varias disciplinas y ciencias, realizados con diferentes objetivos y en diversos contextos, mostró que disminuir de cuatro a tres las opciones de respuesta en los exámenes reduce el nivel de dificultad de los ítems solo en 0,04, incrementa el poder de discriminación en 0,03 y la confiabilidad en 0,02 (5). Igualmente, en los estudios que limitaron la comparación a pruebas con cinco o cuatro opciones, recomiendan las de cuatro. El autor concluye que las pruebas con tres opciones son óptimas para la mayoría de los escenarios (5). Incluso con enfoques teóricos con demostraciones matemáticas, diferentes autores coinciden en que tres es el número óptimo de opciones de respuesta para un examen de selección múltiple (19).

Los enfoques empíricos para el análisis del número adecuado de opciones de respuesta han eliminado el ítem incorrecto con menor discriminación, el menos seleccionado o simplemente por azar e igualmente esta opción se ha reemplazado de diferentes maneras, por azar o por asignación. En general, los autores concluyen que no se presentan cambios en la discriminación o en la confiabilidad o son insignificantes desde el punto de vista práctico (19). Nuestro estudio, llevado a cabo en un campo educativo donde se ha explorado muy poco este aspecto, la medicina, está en la misma dirección de estos resultados, con corroboración de los hallazgos con dos enfoques teóricos diferentes de la medición.

Existen argumentos prácticos para respaldar pruebas con menos opciones de respuesta, algunos de los cuales no se han evaluado objetivamente, pero parecen plausibles. Los profesores generalmente le invierten mucho tiempo al tallo y descuidan la calidad de las opciones de respuesta con lo cual, muy posiblemente, a mayor número de dichas opciones, más probabilidad hay de cometer errores en la elaboración de las mismos; por lo tanto, al disminuirlas se aumentan la calidad y la validez de la pregunta; las pruebas con menos opciones de respuesta pueden incluir mayor número de preguntas en el mismo tiempo que aquellas más cortas con más opciones, con lo cual automáticamente se incrementa la confiabilidad de la prueba (20), se pueden abarcar más temas en la evaluación, se disminuyen los costos de impresión si no se va a aumentar el número de preguntas, hay menos distracción para los evaluados y se van a sentir menos

presionados o van a tener más tiempo y habrá menos oportunidad de aportar claves para algunas respuestas (19).La reducción del número de opciones de respuesta mejora la eficiencia en el uso del tiempo tanto para el profesor como para el evaluado.

Aun con estos resultados, hay cierta resistencia de las instituciones y de los docentes a usar preguntas con solo tres opciones de respuesta y continúan con las de cuatro o incluso cinco, posiblemente por el peso que tiene la tradición en el proceso evaluativo (6). Por otro lado, como llaman la atención algunos autores, posiblemente más importante que el número sea la calidad de las opciones de respuesta y reconocen que no existe respaldo psicométrico para obligar a que todas las preguntas tengan el mismo número de opciones de respuesta, porque de manera natural una puede tener más o menos opciones que parezcan lógicas (6,21), lo cual puede depender también del propósito de la prueba (20). En este sentido, una recomendación es escribir de entrada tantas opciones de respuesta que suenen razonables como sea posible y en un segundo paso un comité evaluador elige las opciones más apropiadas (8,22). Las pruebas bien elaboradas, con buenas opciones de respuesta, son las mejores, independientemente del número de tales opciones. Para lograr esto, se deben buscar opciones de respuesta bien elaboradas, centradas en temas relevantes y revisadas por expertos.

Aunque no fue objetivo de este estudio evaluar la prueba en sí, los valores alfa de Cronbach obtenidos ameritan al menos un comentario. Este índice no mide una propiedad inherente de la prueba, sino una propiedad conjunta entre la prueba y los evaluados y su valor es menor cuanto más homogénea sea la población estudiada. Adicionalmente, aunque se trata de una prueba de medicina, en ella se evalúan múltiples especialidades y se incluyen diferentes competencias de la misma, como el diagnóstico, el tratamiento, la rehabilitación y algunos aspectos teóricos no aplicados, lo cual puede explicar que no se obtuvieran los valores altos esperados tradicionalmente. De todas maneras, el objetivo de esta prueba era la evaluación del cambio, y aunque hubo una disminución con la prueba de tres opciones, este cambio fue poco significativo.

El presente estudio tiene varias fortalezas: el tamaño grande de la muestra, la representatividad amplia de los evaluados, la utilización de dos enfoques teóricos de medición y la concordancia con estudios similares en otras áreas. Una posible debilidad es la forma en que se obtuvo la prueba con tres opciones de respuesta, mediante un criterio estadístico para eliminar la opción menos elegida, y la asignación por azar de la posible respuesta del evaluado a esta pregunta. Este es uno de los métodos utilizados en la literatura, pero asume que el evaluado está eligiendo sus respuestas por azar; sin embargo, se sabe que ante el desconocimiento de una de ellas, los evaluados tienen mecanismos alternativos de elegir la mejor respuesta, buscando claves en el contenido o en la estructura de la pregunta. Dado que para la elaboración final de estas preguntas se tienen en cuenta todas las recomendaciones para evitar que esta misma situación suceda, es poco probable que en la vida real el evaluado pudiera encontrar claves en esa estructura. No es viable aplicar la misma prueba a los mismos evaluados con tres y cuatro opciones, porque para la segunda oportunidad ya las condiciones no serían las mismas; por lo tanto, se considera que esta metodología es válida para la pregunta.

CONCLUSIÓN

Con esta demostración objetiva, se recomienda que se pase a preguntas con solo tres opciones de respuesta, que se utilicen las de cuatro opciones solo cuando por consenso se acuerde que todas parecen razonables y que incluso se baje a dos cuando se considere que un concepto requiere evaluación, pero no es posible redactar más opciones de respuesta que parezcan lógicas. Igualmente, se recomienda que se replantee la directriz rígida de tener un número fijo de opciones de respuesta en los exámenes y que estos dependan de las particularidades de cada pregunta.

CONFLICTO DE INTERÉS

Ninguno que declarar.

REFERENCIAS BIBLIOGRÁFICAS

1. Gimeno Sacristán J. La evaluación en la enseñanza. En: Sacristán Gimeno J, Gómez Pérez AI, editores.

Comprender y transformar la enseñanza. 5ª ed. Madrid: Morata; 1996. p. 334-97.

2. Haladyna TM, Downing SM, Rodriguez MC. A Review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002 Jul;15(3):309-33.

3. García-Garro AJ, Ramos-Ortega G, Díaz de León-Ponce MA, Olvera-Chávez A. Instrumentos de evaluación. Rev Mex Anestesiol. 2007;30(3):158-64.

4. Moreno R, Martínez RJ, Muñiz J. Directrices para la construcción de ítems de elección múltiple. Psicothema. 2004;16(3):490-7.

5. Rodriguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. Educ Meas Issues Pract. 2005 Jun;24(2):3-13.

6. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ. 2009 Jul;9:40.

7. Vyas R, Supe A. Multiple choice questions: a literature review on the optimal number of options. Natl Med J India. 2008;21(3):130-3.

8. 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 Med Educ. 2010 Jan;10(1):85.

9. Kasule OH. Overview of medical student assessment: Why, what, who, and how. J Taibah Univ Med Sci. 2013 Aug;8(2):72-9.

10. Morrison S, Free KW. Writing multiple-choice test items that promote and measure critical thinking. J Nurs Educ. 2001 Jan;40(1):17-24.

11. Brady AM. Assessment of learning with multiple-choice questions. Nurse Educ Pract. 2005 Jul;5(4):238-42.

12. McMahan CA, Pinckard RN, Prihoda TJ, Hendricson WD, Jones AC. Improving multiple-choice questions to better assess dental student knowledge: distractor utilization in oral and maxillofacial pathology course examinations. J Dent Educ. 2013 Dec;77(12):1593-609.

13. Matlock-Hetzel S. Basic Concepts in Item and Test Analysis [Internet]. En: Annual meeting of the Southwest Educational Research Association; Austin, January, 1997. Texas: Texas A&M University; 1997 [consultado 2014 Abr 15]. Disponible en: http://ericae.net/ft/tamu/Espy.htm

14. Backhoff Escudero E, Larrazolo Reyna N, Rosas Morales M. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). REDIE. 2000;2(1):12-29.

15. Mitra NK, Nagaraja HS, Ponnudurai G, Judson JP. The levels of difficulty and discrimination indices in type a multiple choice questions of pre-clinical semester 1, multidisciplinary summative tests. IeJSME. 2009;3(1):2-7.

16. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004 Sep;38(9):1006-12.

17. Subkoviak MJ. A Practitioner's Guide to computation and interpretation of reliability indices for mastery tests. J Educ Meas. 1988 Mar;25(1):47-55.

18. Altman DG, Bland JM. Measurement in Medicine : the analysis of method comparison studies. Statistician. 1983 Sep;32(3):307-17.

19. Shizuka T, Takeuchi O, Yashima T, Yoshizawa K. A comparison of three-and four-option English tests for university entrance selection purposes in Japan. Lang Test. 2006 Jan;23(1):35-57.

20. Baghaei P, Amrahi N. The effects of the number of options on the psychometric characteristics of multiple choice items. Psychol Test Assess Model. 2011;53(2):192-211.

21. Frary RB. More multiple-choice item writing do´s and dont´s. Pract Assess Res Eval. 1995;4(11):1-6.

22. Swanson DB, Holtzman KZ, Allbee K. Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. Acad Med. 2008 Oct;83(10 Suppl):S21-4.