Artículo original
Med Int Méx 2015;31:259-273.

Modificación de los índices psicométricos de los exámenes departamentales diseñados para la evaluación de los residentes de Medicina Interna del posgrado de la Facultad de Medicina, UNAM, en función de cinco o cuatro opciones de respuesta

Perla Patricia Borrego-Mora[1]
Marco Antonio Santana-Borrego[2]

[1] Departamento de Evaluación, Posgrado, Facultad de Medicina, UNAM.
[2] PTA Sistemas, empresa de desarrollo de software.

RESUMEN

Antecedentes: el número de opciones se ha discutido ampliamente entre los partidarios de los ítems de opción múltiple. Se puede elegir la confiabilidad, validez o ambas como criterio para establecer un número de alternativas de respuesta.

Objetivo: analizar si al disminuir de cinco a cuatro opciones de respuesta se modifican las propiedades psicométricas de los exámenes departamentales diseñados para la evaluación de los residentes de especialidades médicas y, por ende, la calidad del instrumento de evaluación.

Material y método: estudio comparativo cuasiexperimental en el que se aplicó un examen con 300 preguntas de opción múltiple en dos versiones: cinco y cuatro opciones de respuesta a dos grupos de 117 residentes de Medicina Interna cada uno.

Resultados: se obtuvo un coeficiente de correlación de 0.925 ($p<0.05$), el coeficiente de confiabilidad con IC a 95% para cinco opciones fue de 0.913 y para cuatro opciones fue de 0.905. El índice de dificultad para cinco opciones fue de 63 y para cuatro fue de 65; 54% de los reactivos tuvo mejor índice de dificultad en la versión de cinco opciones. La correlación punto biserial para cinco y cuatro opciones fue de 0.16 y 0.18, respectivamente.

Conclusiones: en la versión con cuatro opciones, 54% de los reactivos los contestó un mayor número de residentes, incluidos los de bajo rendimiento, éstos los pudieron acertar al azar (20% de probabilidad), lo que permite que residentes limítrofes obtengan calificación aprobatoria. Al comparar ambas versiones, con cuatro opciones resulta 44% de reactivos con índice de discriminación menor, lo que indica que esa versión tiene menor capacidad de discriminar entre los residentes de puntuación alta y baja. Al observar estas modificaciones consideramos que cinco opciones son mejores que cuatro para la evaluación de los residentes de especialidades médicas.

Palabras clave: preguntas de opción múltiple, cinco y cuatro opciones de respuesta, índice de discriminación, índice de confiabilidad, índice de dificultad, índices psicométricos.

Recibido: 29 de enero 2015

Aceptado: 14 de abril 2015

Correspondencia: Perla Patricia Borrego Mora
Subdivisión de Estudios de Posgrado
Facultad de Medicina
Universidad Nacional Autónoma de México
Av. Universidad 3000
0410 México, DF
patriciaborregomora@hotmail.com

Este artículo debe citarse como
Borrego-Mora PP, Santana-Borrego MA. Modificación de los índices psicométricos de los exámenes departamentales diseñados para la evaluación de los residentes de Medicina Interna del posgrado de la Facultad de Medicina, UNAM, en función de cinco o cuatro opciones de respuesta. Med Int Méx 2015;31:259-273.

Modification of psychometric indexes of departmental examinations designed to assess residents of Internal Medicine, Graduate School of Medicine, UNAM, according to five or four answer choices

ABSTRACT

Background: The number of options has been widely discussed among the supporters of the multiple-choice items. It is possible to choose the reliability, validity, or both as a criterion for establishing a number of possible answers.

Objective: To analyze if reducing for five to four options of answer modifies psychometric properties of departmental exams designed to assess medical specialties residents and, thus, the quality of assessment instrument.

Material and method: A comparative quasi-experimental study was done applying a test with 300 multiple choice questions in two versions: 5 and 4 possible answers to two groups of 117 internal medicine residents, each one.

Results: We obtained a correlation coefficient of 0.925 ($p<0.05$), the coefficient of reliability with 95% CI for five answers was 0.913 and for four options was 0.905. The index of difficulty for five options was 63 and for four was 65, 54% of the items had a better index of difficulty in the version of five options. The point biserial correlation for five and four options was 0.16 and 0.18, respectively.

Conclusions: In the version with four options, 54% of the items answered a greater number of residents, including low achievers; they could hit the random (20% probability), allowing bordering residents to obtain a passing grade. By comparing the two versions, with four options 44% were reactive with a lower discrimination index, which indicates that they were less able to discriminate between residents of high and low score. By observing these changes we believe that five options are better than four to assess medical specialties.

Key words: multiple choice questions, five and four response options, discrimination index, reliability index, difficulty index, psychometric indexes.

ANTECEDENTES

Las pruebas son un conjunto homogéneo y estandarizado de ítems cuyo objetivo es la evaluación cuantitativa en condiciones rigurosamente estandarizadas de rasgos y atributos psicológicos y educacionales. Cronbach sugiere que las pruebas se refieren a "…procedimiento sistemático para observar la conducta y describirla con la ayuda de escalas numéricas o

categorías establecidas […], el evaluador recoge información preguntando y observando a todas las personas de la misma manera, en la misma o en comparables situaciones".[1]

Las pruebas constan de preguntas, tareas, estímulos, situaciones, etcétera, que intentan poner en relieve una muestra de las conductas del sujeto, representativa de la característica que se quiere apreciar o medir; contiene preguntas estructuradas destinadas a evaluar la frecuencia o intensidad en que se manifiesta determinado atributo o rasgo (dominio del conocimiento).[2]

Entre las pruebas, las construidas en opción múltiple son las más recomendadas y útiles; éstas se han centrado en una serie de reglas metodológicas en cuanto al formato y escritura, entre las que se encuentra la redacción de las opciones de respuesta, que deben parecer razonables a los estudiantes que tienen menos conocimientos que la mayoría de sus compañeros de clase.[3,4]

Cada uno de los distractores debe diferir de la opción correcta en un solo aspecto; de todas las opciones, la correcta es la única que reúne todos los elementos apropiados o definitorios. Un buen distractor es el que se aproxima inexactamente a la opción correcta. Si los distractores difieren demasiado de la opción correcta, éstos dejan de ser plausibles y el reactivo pierde validez, así que deben producirse distractores que se alejen un poco, aunque no demasiado, de la opción correcta.

Según Ebel, la producción sistemática de errores puede ayudar a clasificar todos los errores posibles en que se puede incurrir en un procedimiento. Esto es muy útil para el análisis de errores en el aprendizaje de ciertas habilidades.[5]

El número de opciones es quizá lo discutido más intensamente entre los partidarios de los ítems de opción múltiple. Ha habido discusiones fuertes en cuanto a tres, cuatro y cinco opciones. Los que están contra cinco opciones creen que tres e incluso cuatro opciones aumentan la probabilidad de que un estudiante adivine la respuesta correcta. Los que están a favor de tres opciones insisten que sus pruebas pueden ser tan eficaces como una de cuatro y cinco opciones, porque los distractores adicionales probablemente son menos factibles.

Una vez que se decide el número de opciones deseadas, es recomendable utilizar todos los ítems con este número de opciones en todo el examen para reducir la posibilidad de errores.[3]

Los primeros estudios del número de opciones de respuesta en una prueba se remontan hasta el primer tercio del siglo XX y se centran principalmente en pruebas diseñadas para evaluar el rendimiento académico. De acuerdo con Lord, entre los primeros en investigar este tema está Toops. En general, los autores coinciden en señalar que tres es el número óptimo de alternativas de respuesta porque permite garantizar la adecuada fiabilidad de los instrumentos que evalúan el rendimiento académico.[6,7] Los estudios empíricos posteriores parecen confirmar esta conclusión,[8-10] que tiene el respaldo desde la perspectiva de la Teoría Clásica de los Tests[2,9-15] y de la Teoría de Respuesta a los Ítems.[16,17] En la misma línea, Rogers y Harley[15] señalan que para este tipo de instrumentos tres opciones de respuesta son tan buenas como cuatro o cinco en términos de discriminación del ítem, de la consistencia interna y del error estándar.[18-23]

Otros autores sostienen que basta con utilizar dos o tres alternativas de respuesta, otros sugieren utilizar desde cinco a siete opciones hasta un máximo de 25 alternativas de respuesta para maximizar las propiedades psicométricas del instrumento.[24-30]

Existen discrepancias respecto de qué criterio utilizar como mejor indicador para establecer un número óptimo de alternativas de respuesta. En general, la mayoría de los investigadores suelen elegir la confiabilidad como criterio.[20,31-35] Otros, aunque en menor medida, recurren a la validez[36,37] o simultáneamente a estos dos criterios.[21,23,26,38,39]

Este problema puede tratarse desde diferentes modelos psicométricos, como la Teoría Clásica de los Tests,[20,30-32,35] el Análisis Factorial de los Ítems[36-39] o la Teoría de Respuesta a los Ítems.[25,40]

La Teoría Clásica de los Tests es el primer modelo propuesto para abordar aspectos relacionados con la confiabilidad del instrumento de medición. Surgió a comienzos del siglo XX, con la formulación que hace Spearman[40-44] de lo que hoy se conoce como Modelo de la Puntuación Verdadera o TCT. Posteriormente, Lord y Novick publicaron el libro *Statistical theories of mental tests scores*, en el que se plantea una revisión crítica de la Teoría Clásica de los Tests, al tiempo que proponen nuevas líneas de trabajo, entre las que se encuentra la Teoría de Respuesta a los Ítems.

La Teoría Clásica de los Tests se encarga de establecer un modelo capaz de evaluar las propiedades psicométricas de los instrumentos de evaluación. Específicamente, estudia los factores que influyen en la puntuación observada en las pruebas y propone modelos que permiten estimar la puntuación verdadera obtenida por el sujeto, esto último a partir de inferencias basadas en las puntuaciones observadas.[3,4,7]

Por tanto, consideramos tres indicadores matemáticos que se constituyen en parámetros de cada pregunta de opción múltiple: confiabilidad, dificultad y discriminación.

Confiabilidad

Es una estimación del grado de consistencia o constancia entre mediciones repetidas efectuadas a los sujetos con el mismo instrumento.

Debido a que en el campo de las ciencias de la conducta es imposible obtener el mismo resultado utilizando el mismo instrumento (someter a un mismo grupo de alumnos a una misma prueba producirá seguramente resultados diferentes, porque los individuos no se comportan dos veces de manera idéntica), los principios en los se apoya el concepto de confiabilidad de una prueba son:

a) El número de ítems incluidos en la prueba. Si el número de reactivos es bajo, hay más posibilidades de que los resultados no sean confiables. Un mayor número de reactivos, en general, posee una muestra mayor de lo que se intenta medir y tiende a cancelar los efectos del azar. La tabla de Ebel ilustra esta relación entre la longitud de la prueba y su confiabilidad (Cuadro 1).[5]

Se observa que mientras un examen con cinco reactivos produce confiabilidad de 0.20, uno con 640 la da de 0.97, pero el aumento indiscriminado del número de reactivos influye en la adecuación de la prueba y en el tiempo disponible para su resolución.

Cuadro 1. Relación entre el número de reactivos e índice de confiabilidad, tabla de Ebel

Núm. de reactivos	Confiabilidad
5	0.20
10	0.33
20	0.50
40	0.66
80	0.80
160	0.89
320	0.94
640	0.97
> 640	0.99

b) Grado de homogeneidad de los elementos. Las pruebas que traten de comprobar conocimientos o información de un solo tema específico tenderán a ser más confiables que las pruebas globales. Si la decisión del educador es incluir distintos tipos de temas de diversa índole, porque esto conviene a los fines de la evaluación (un examen promocional por ejemplo), debe tener en cuenta que la confiabilidad de los resultados se ve necesariamente disminuida y, por tanto, deberá trabajar en las otras características que tienden a aumentar el grado de confiabilidad.

Índice de confiabilidad (IC) alfa de Cronbach. Alfa depende de la función de dos componentes de la prueba: el número de ítems (o longitud de la prueba) y la proporción de variancia total de la prueba debida a la covariancia entre sus partes (ítems). Ello significa que la fiabilidad de la prueba depende de su longitud y de la covariancia entre sus ítems. Alfa se expresa como: prueba total. Este coeficiente es la forma más habitual de estimación de fiabilidad cuando se aplica el método de consistencia interna en pruebas basadas en la Teoría Clásica de los Tests y se basa en el siguiente modelo matemático (Modelo 1):

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i}^{K=1} \sigma^2 Y_i}{\sigma^2 x}\right)$$

donde:

α: índice de confiabilidad.

Σ: sumatoria de las variancias de los ítems.

σ^2: variancia de la suma de los ítems.

K: número de ítems.

Dificultad

Es el porcentaje de sustentantes que contestan correctamente una pregunta; lo que permite catalogarlas desde muy fácil hasta muy difícil. Una prueba que posea dificultad media es superior en confiabilidad de resultados a las que tienen mucha dificultad y a las que resultan muy fáciles. Un buen examen será el que sea superado por algo más de la mitad de los educandos.

En otras palabras, la sucesiva corrección del grado de dificultad de cada pregunta según el porcentaje de alumnos que la respondieron bien en cada uno de los exámenes en los que se la incluyó hace que el educador cuente con un parámetro de dificultad histórico de cada pregunta en sí y, por tanto, incluirá en una prueba el conjunto de preguntas que ya ha mostrado tener dificultad media, independientemente del resultado que arroje el examen conformado en el grupo particular de educandos a los que se examinará.[45]

Discriminación

El grado de discriminación de una pregunta es la capacidad del reactivo de distinguir o discriminar entre los alumnos que dominan la materia sujeta a examen y los que no la dominan (capacidad de separar a los alumnos entre los bien preparados y los mal preparados para el examen); es un parámetro de fundamental importancia.

Una prueba que contenga reactivos con un promedio alto de grado de discriminación logra resultados más confiables que otra que incluya reactivos con menor promedio de ese grado. Si el docente tiene una herramienta informática que evalúe el grado de discriminación de cada pregunta almacenada en su banco, podrá recurrir a aquéllas con mayor grado de discriminación en las pruebas finales, a fin de contrarrestar el efecto de la alta homogeneidad de los reactivos que ese tipo de exámenes debe incluir.

Un cálculo matemático sencillo del grado de discriminación que es de rápida obtención y, por tanto, resulta práctico es el Índice de Pemberton,

que consiste en dividir el grupo de examinados en tres: 27% de mayor puntaje, 27% de menor puntaje y un tercer grupo intermedio. Para calcular el índice se descarta el grupo de resultados intermedios y se resta el número de respuestas correctas de ambos grupos, dividiendo el resultado por el número total de alumnos sometidos al examen, según el siguiente modelo matemático (Modelo 2):

$$Di = \frac{GA_{aciertos} - GB_{aciertos}}{N_{grupo\ mayor}}$$

donde:

Di: índice de discriminación del reactivo *i*.

$GA_{aciertos}$: número de aciertos en el reactivo *i* de 27% de personas con las puntuaciones más altas en la prueba.

$GB_{aciertos}$: número de aciertos en el reactivo *i* de 27% de personas con las puntuaciones más bajas en la prueba.

$N_{grupo\ mayor}$: número de personas en el grupo más numeroso (GA o GB).

El índice de Pemberton tiene su justificación teórica en el hecho de que un sustentante dé respuesta correcta o incorrecta a un reactivo, sólo provee información para la determinación del grado de dificultad que él mismo encontró en la pregunta, pero nada nos dice del reactivo en sí. El procedimiento correcto para la determinación del grado de discriminación de una pregunta consiste en relacionar la respuesta de un reactivo en particular (correcta o incorrecta) con el resultado del examen de cada sustentante. En otras palabras, la pregunta discriminará bien si es contestada bien por los sustentantes que pasaron satisfactoriamente el examen y mal por los que fracasaron en el examen.

El índice de discriminación de un reactivo debe disminuir si el mismo fue contestado bien por un alumno del grupo inferior, también debe disminuir si fue contestado mal por un alumno del grupo superior. En consecuencia, el índice aumenta si el reactivo es contestado bien por un alumno del grupo superior y mal por un alumno del grupo inferior (Cuadro 2).

Cuadro 2. Poder de discriminación de los reactivos según su discriminación

Discriminación	Calidad	Recomendaciones
> 0.39	Excelente	Conservar
0.30-0.39	Buena	Posibilidades de mejorar
0.20-0.29	Regular	Necesidad de revisar
0.10-0.19	Mala	Revisar a profundidad o descartar
< 0.10	Pésima	Descartar

Dos indicadores más de la efectividad discriminativa de un reactivo son el punto de correlación biserial y el coeficiente de discriminación. La ventaja de utilizar el coeficiente de discriminación en lugar del índice de discriminación (ID) es que con el primer método se toman en cuenta todas y cada una de las personas evaluadas, mientras que con el segundo, sólo se toma 54% de ellas (27% más alto y 27% más bajo). Además, es un coeficiente de correlación entre las puntuaciones en el ítem y la puntuación global de la prueba. Por tanto, puede ser más sensible para detectar el comportamiento de los ítems, según el siguiente modelo matemático (Modelo 3):

$$CD = \frac{\sum(xy)}{n s_x s_y}$$

donde:

CD: coeficiente de discriminación.

$\sum(xy)$: suma de los productos de las desviaciones de las puntuaciones en el ítem y las puntuaciones en toda la prueba.

n: número de respuestas dadas a esta pregunta.

s_x: desviación típica de las puntuaciones fraccionales para esta pregunta.

s_y: desviación típica de las puntuaciones en todo el cuestionario.

Este parámetro adopta valores entre +1 y -1, los valores positivos indican los ítems que discriminan entre estudiantes competentes y no competentes, en tanto que los valores negativos se dan cuando los ítems son mejor contestados por los estudiantes con calificaciones más bajas.

Coeficiente de correlación biserial (r_{bis})

Se calcula para determinar el grado en que las competencias que mide la prueba también las mide el ítem. El r_{bis} proporciona un cálculo de la correlación producto-momento de Pearson entre la calificación total de la prueba y el continuo hipotético del reactivo, cuando éste se dicotomiza en respuestas correctas e incorrectas.

Correlación del punto biserial (r_{pbis})

Se utiliza para saber si las personas "adecuadas" son las que obtienen las respuestas correctas, qué tanto poder predictivo tiene el reactivo y cómo puede contribuir a las predicciones.

El r_{pbis} nos dice más acerca de la validez predictiva de la prueba que el r_{bis}, porque éste tiende a favorecer los reactivos de dificultad media. También se sugiere que el r_{pbis} es una medida que combina la relación entre el criterio del reactivo y el nivel de dificultad, según el siguiente modelo matemático (Modelo 4):

$$r_{pbis} = \frac{\overline{\chi}_1 - \overline{\chi}_0}{s_\chi} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

donde:

x_1: media de las puntuaciones totales de los que respondieron correcto un ítem.

x_0: media de las puntuaciones totales de los que respondieron incorrecto un ítem.

S_x: desviación estándar de las puntuaciones totales.

n_1: número de sustentantes que respondieron correcto el ítem.

n_0: número de sustentantes que respondieron incorrecto el ítem.

Coeficiente de correlación de Pearson

Mide la relación lineal entre dos variables aleatorias cuantitativas, el valor de correlación varía en el intervalo [-1, +1].

Si r=1, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables, denominada relación directa.

Si $0 < r < 1$, existe una correlación positiva.

Si r=0, no existe relación lineal entre las dos variables.

Si $-1 < r < 0$, existe una correlación negativa.

Si r=-1, existe una correlación negativa perfecta.

Nuestra hipótesis es que cuando se incluyen cuatro o cinco opciones de respuesta se modifican las propiedades psicométricas de los exámenes departamentales de las especialidades médicas de la Facultad de Medicina de la UNAM.

El objetivo de este artículo es analizar si al disminuir de cinco a cuatro opciones de respuesta se modifican las propiedades psicométricas de los exámenes departamentales diseñados para la evaluación de los residentes de las especialidades médicas y, por ende, la calidad del instrumento de evaluación.

MATERIAL Y MÉTODO

Estudio comparativo cuasiexperimental efectuado en dos grupos de residentes de la especialidad de Medicina Interna a los que se aplicaron dos versiones de examen; una versión a cada grupo.

La variable independiente fue el número de opciones de respuesta. Las variables dependientes fueron: índice de dificultad, de confiabilidad y de discriminación, correlación punto biserial y coeficiente de correlación.

Definición de variables

Número de posibilidades de contestar una pregunta: cuatro y cinco opciones.

Porcentaje de sustentantes que contestan correctamente cada pregunta de cada versión del examen; con valores entre 0 y 100%.

Índice de confiabilidad de cada versión del examen, calculado mediante el modelo matemático 1; alfa de Cronbach con valor de 0 a 1.

Índice de discriminación de cada pregunta de cada versión del examen, calculado mediante el modelo matemático 2; con valor de -1 a 1.

Correlación punto biserial de cada versión del examen, calculado mediante el modelo matemático 4 con valor de -1 a 1.

Coeficiente de correlación entre las versiones del examen, calculado mediante modelo matemático con valor de -1 a 1.

Los profesores titulares y adjuntos de la especialidad de Medicina Interna de los 25 cursos reconocidos por la UNAM elaboraron para los residentes de segundo año un examen estructurado con 60 casos clínicos que incluyó 300 preguntas en la técnica de opción múltiple con una sola respuesta correcta, con cinco opciones posibles de respuesta (versión de cinco opciones); posteriormente se les pidió en una reunión conjunta que eliminaran la opción de respuesta menos factible de ser contestada y así se generó una versión con los mismos casos clínicos y las mismas preguntas, pero con cuatro opciones posibles de respuesta (versión de cuatro opciones).

El examen departamental en sus dos versiones se aplicó en las instalaciones de la Facultad de Medicina, en día sábado, a las 8:00 horas con duración de tres horas.

Los 234 alumnos inscritos en el segundo año del plan de estudios fueron listados por orden alfabético de su primer apellido; a los alumnos en posición non se le aplicó la versión de cinco opciones y a los alumnos en posición par, la versión de cuatro opciones de respuesta.

El análisis de los datos se hizo a través de los paquetes estadísticos MICROCAT Testing System 3.5 y SPSS 11.0.

RESULTADOS

En la versión de cinco opciones se obtuvo un promedio de 191.17 y para cuatro opciones de 196.4, con desviación estándar de 24.5 y 23.19, respectivamente; ambas versiones resultaron multimodales. Se observó una diferencia de 15 puntos en la puntuación mínima entre la versión de cuatro opciones respecto de la de cinco; y para la puntuación máxima fueron cinco puntos de diferencia entre la versión de cuatro respecto de la de cinco opciones (Cuadro 3).

Cuadro 3. Descripción estadística

	Cinco opciones	Cuatro opciones
Promedio	191.17	196.48
Moda	186, 204	174, 193, 200, 208
Mediana	193	198
Desviación estándar	24.56	23.19
Puntuación mínima	125	140
Puntuación máxima	246	251
Confiabilidad (alfa Crombach)	0.913	0.905
Porcentaje promedio	63.7	65.5

Se observa que el número de residentes que se encuentran alrededor de promedio ±0.5 de desviación estándar es mayor en la versión de cinco que en la de cuatro opciones (44 contra 39). Cuando al promedio se le restan una y dos desviaciones se observa menor número de residentes en la versión de cinco que en la de cuatro opciones (26 contra 33), lo que se traduce en que los residentes de bajo rendimiento, al darles cuatro opciones de respuesta, obtienen mejor puntuación que cuando se les dan cinco opciones, lo que se puede explicar por acertar al contestar al azar. En contrapartida, cuando al promedio se le suman 1 y 2 desviaciones estándar se observa un número semejante de residentes en ambas versiones (36 contra 33), lo que se traduce en que los residentes de alto rendimiento, al darles cinco opciones de respuesta, obtienen puntuación semejante que cuando se les dan cuatro opciones, lo que se puede explicar porque un residente de alto rendimiento intenta responder una pregunta a través de la eliminación de opciones (exclusión) y no por azar.

No se observó diferencia entre ambas versiones del examen en relación con el número de residentes que obtuvieron menos de 180 aciertos (60%) del examen, así como en el número de alumnos por debajo del promedio (Cuadro 4 y Figuras 1 y 2).

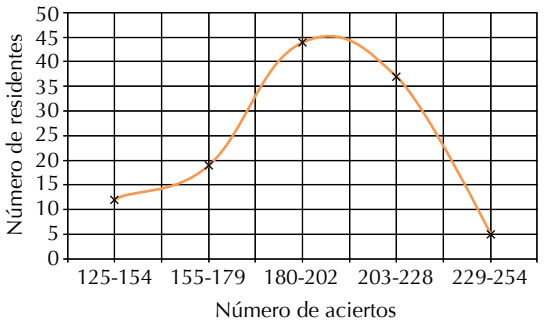

Figura 1. Distribución de aciertos en la versión de cinco opciones.

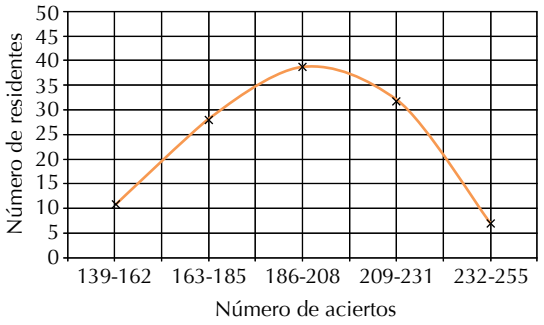

Figura 2. Distribución de aciertos en la versión de cuatro opciones.

Cuadro 4. Distribución de aciertos de acuerdo con el promedio y desviación estándar (DS)

	Cinco opciones			Cuatro opciones		
	Aciertos	**Núm. de alumnos**	**Porcentaje de alumnos**	**Aciertos**	**Núm. de alumnos**	**Porcentaje de alumnos**
Menos 2 DS	125-154	12	10	139-162	11	9
Menos 1 DS	155-179	19	16	163-185	28	24
Promedio ±0.5 DS	180-202	44	38	186-208	39	33
Más 1 DS	203-228	37	32	209-231	32	27
Más 2 DS	229-254	5	4	232-255	7	6

Coeficiente de correlación

El valor de la correlación entre las dos versiones del examen fue 0.925 (con significación de 0.05), por lo que se asume que existe una fuerte relación lineal entre ambos instrumentos.

Una vez ajustada la recta de regresión a la nube de observaciones, fue importante medir la bondad del ajuste realizado, lo que permitió decidir que el ajuste lineal fue suficiente, para lo que se utilizó el coeficiente de determinación r^2 (cuadrado del coeficiente de correlación; valores de referencia entre 0 y 1, entre menor sea, indica mayor variabilidad), que resultó con valor de 0.85, lo que explicó la variación entre la versión de cuatro opciones de respuesta respecto de la de cinco opciones (Figura 3).

Índice de confiabilidad alfa de Cronbach

Se evaluó el efecto de variar el número de opciones de respuesta de cinco a cuatro opciones en la consistencia interna, de acuerdo con el modelo de la Teoría Clásica de los Tests.

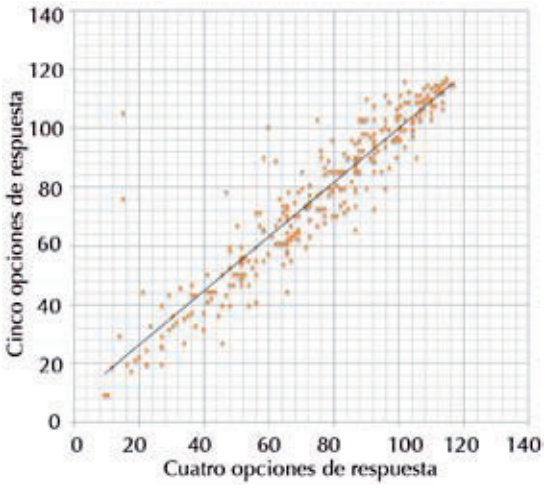

Figura 3. Correlación entre ambas versiones de examen.

Se obtuvieron los intervalos de confianza a 95% de los coeficientes alfa de cada versión, para comparar la significación estadística de las diferencias observadas entre ambos coeficientes alfa.

Para la versión de cinco opciones se obtuvo un alfa de 0.913, con intervalos de confianza a 95% de 0.879, 0.941 y para la de cuatro opciones alfa de 0.905 e intervalos de confianza de 0.871, 0.933.

Se observa que el valor de alfa para la versión de cuatro opciones se encuentra en los límites del intervalo de confianza de la versión con cinco opciones; dar cuatro opciones de respuesta no difiere significativamente con la versión de cinco opciones de respuesta. Se concluye que no existe efecto del número de opciones de respuesta en la consistencia interna del examen entre las versiones del examen por tener intervalos de confianza solapados.

Índice de dificultad (*p*)

Al analizar los grados de dificultad de cada versión del examen, podemos señalar que el valor promedio de *p*, para la versión de cinco opciones fue de 63 y para la de cuatro fue de 65, lo que indica que los reactivos del examen con cinco opciones los contesta un porcentaje menor de residentes (Cuadro 5).

Al realizar el análisis del comportamiento de la dificultad de los 300 reactivos se observa que para la versión de cinco opciones, 33% de los reactivos tuvo dificultad de 0 a 45 (difíciles y muy difíciles), mientras que para la versión de cuatro opciones, 31% tuvo la misma condición; en lo que se refiere a la dificultad media, en la versión de cinco opciones, 33% de los reactivos estuvo entre 46 y 65; mientras que para la versión de cuatro opciones fue de 25%, y en lo que respecta a muy fáciles y fáciles (66 a 100) para la versión de cinco opciones se obtuvo 34% y para la de cuatro fue de 44%.

Cuadro 5. Número de reactivos de acuerdo con su índice de dificultad

Dificultad	Cinco opciones Núm. (%)	Cuatro opciones Núm. (%)
0-5	8 (3)	6 (2)
6-15	14 (5)	14 (5)
16-25	17 (6)	13 (4)
26-35	31 (10)	29 (10)
36-45	28 (9)	30 (10)
46-55	45 (15)	39 (13)
56-65	53 (18)	36 (12)
66-75	45 (15)	49 (16)
76-85	46 (15)	54 (18)
86-95	13 (4)	30 (10)

Se observa que el número de reactivos con dificultad media fue mayor en la versión de cinco opciones que en la de cuatro (33 contra 25), lo que se traduce en mayor número de reactivos con mejor índice de dificultad.

A 124 (41%) reactivos con cinco opciones se les disminuyó el índice de dificultad (lo contestó un porcentaje menor de residentes) al compararlo con el de cuatro opciones, en 14 reactivos (5%) el índice no se modificó (lo contestó igual porcentaje de residentes) y a 162 (54%) se les aumentó el índice de dificultad (lo contestó un porcentaje mayor de residentes). Figura 4

Índice de discriminación

Al analizar el comportamiento del índice de discriminación de los 300 reactivos se observó que para la versión de cinco opciones 18% de reactivos obtuvo discriminación entre 0.36 y 0.65, mientras que para la versión de cuatro opciones fue de 20%; entre 0.16 y 0.35 la versión de cinco opciones tuvo 42% de reactivos y la versión de cuatro opciones, 41%; entre 0 y 0.15 la versión de cinco opciones tuvo 32% y la versión de cuatro opciones, 29%, y de reactivos con discriminación negativa la versión de cinco tuvo 8% y la de cuatro, 10% (lo que indica que mayor número de resistentes con baja puntuación lo contestaron correctamente). Cuadro 6

Al comparar reactivo contra reactivo de la versión con cinco opciones respecto de la de cuatro, se observó que 132 (44%) reactivos tuvieron un índice de discriminación menor (lo contestó un porcentaje mayor de residentes de baja puntuación), lo que permite decir que la versión de cuatro opciones tiene menor capacidad de discriminar entre los residentes de puntuación alta y baja; en 7 (2%) el índice no se modificó (lo contestó igual porcentaje de sustentantes de puntuación alta y baja) y 161 (54%) tuvieron un índice de discriminación mayor (lo contestó un porcentaje mayor de residentes de puntuación alta). Figura 5

Correlación punto biserial

La correlación punto biserial para cinco y cuatro opciones fue de 0.16 y 0.18, respectivamente. El valor de la correlación punto biserial mayor de 0.20 indica que el ítem pertenece a la escala correspondiente, es decir, el ítem mide el mismo rasgo que la escala en conjunto.

CONCLUSIONES

Se obtuvo un coeficiente de correlación de 0.925 ($p<0.05$), lo que indica una fuerte intensidad de asociación entre ambas versiones del examen, lo que significa que con ambas versiones se está explorando un mismo constructo y, de acuerdo con el coeficiente de determinación de 85%, la variabilidad se debe al diferente número de opciones presentadas en las versiones del examen.

En el coeficiente de confiabilidad con intervalos de confianza (IC) a 95% se encontró que el valor de alfa para cuatro opciones está dentro de los intervalos de confianza 0.871 y 0.933 de la versión con cinco opciones, lo que indica que la versión

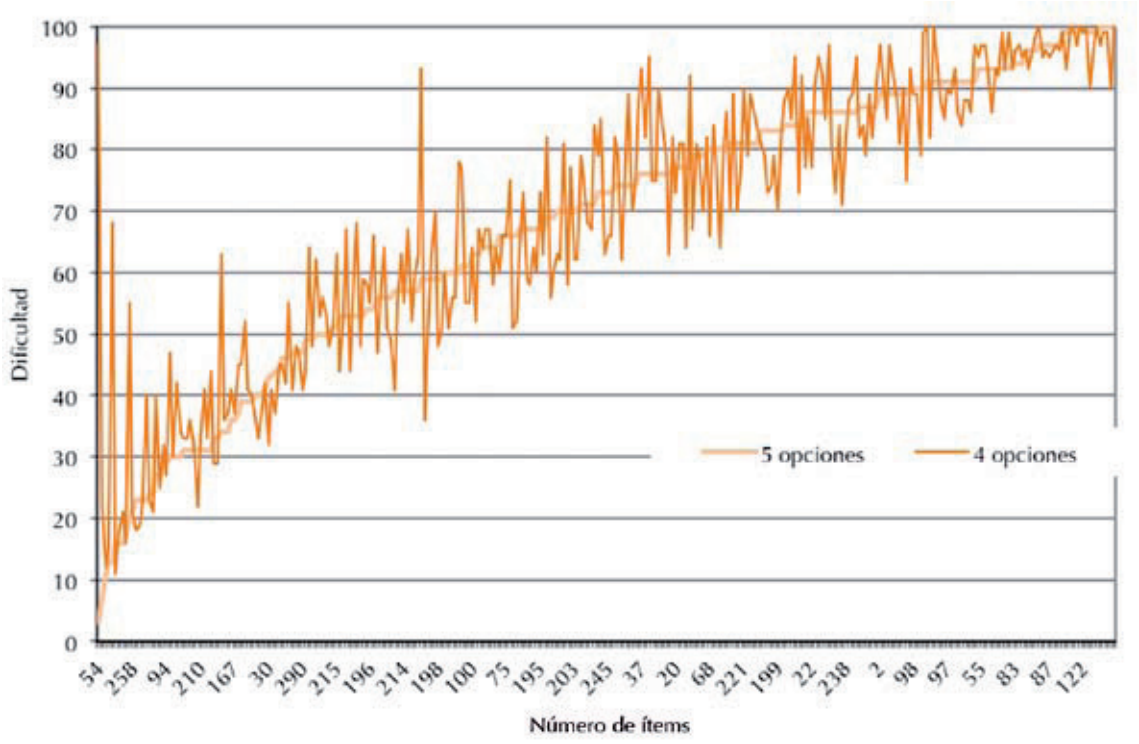

Figura 4. Distribución del índice de dificultad por ítems entre las dos versiones del examen.

Cuadro 6. Distribución de reactivos de acuerdo con su índice de discriminación en la versión de cinco y cuatro opciones de respuesta del examen

Discriminación	Cinco opciones Núm. (%)	Cuatro opciones Núm. (%)
Negativos	25 (8)	31 (10)
0-0-0.05	34 (11)	28 (9)
0.06-0.15	62 (21)	61 (20)
0.16-0.25	62 (21)	64 (21)
0.26-0.35	63 (21)	59 (20)
0.36-0.45	33 (11)	41 (14)
0.46-0.55	13 (4)	14 (5)
0.56-0.65	8 (3)	2 (1)

de cinco opciones no difiere significativamente de la versión de cuatro opciones en cuanto a la consistencia interna del examen. Este índice es adecuado también por tener 300 preguntas.

La media de la dificultad en la versión con cuatro opciones es superior a la de cinco opciones, lo que siginifica que más reactivos fueron contestados por mayor número de residentes, incluidos los que tienen bajo rendimiento, lo que puede explicarse porque estos últimos los pudieron acertar al contestar al azar (20% de probabilidad), lo que permite que residentes limítrofes obtengan una calificación aprobatoria.

Sin embargo, al observar el comportamiento gráfico de cada reactivo en las dos versiones del examen, se observan diferencias entre ambas, en la versión con cinco opciones 41% de los reactivos lo contestó menor número de residentes al

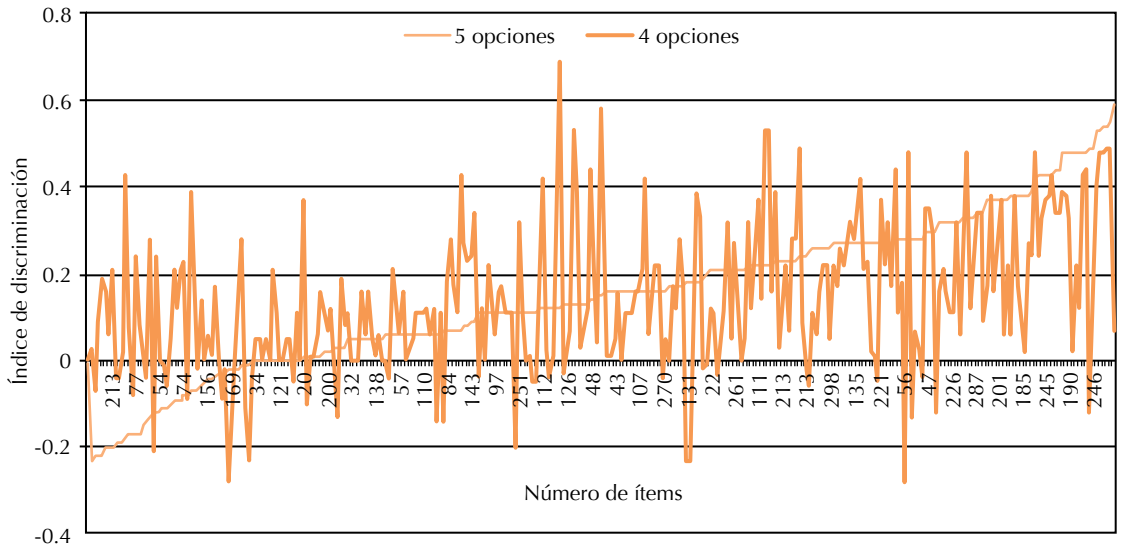

Figura 5. Distribución del índice de discriminación entre las dos versiones del examen.

compararlos con la versión de cuatro opciones y 54% de los reactivos lo contestó mayor número de residentes.

La media de la discriminación para cinco opciones fue de 0.19 y para cuatro opciones de 0.20; para ambas versiones éste es un valor bajo.

Sin embargo, al observar el comportamiento gráfico de las dos versiones del examen se observan diferencias entre ambas; en la versión con cinco opciones 54% de reactivos tuvo mejor índice de discriminación (lo contestó un porcentaje mayor de residentes de puntuación alta) cuando se comparan con cuatro opciones de respuesta y 44% de reactivos tienen menor índice de discriminación (lo contestó un porcentaje mayor de residentes de puntuación baja), cuando se comparan entre cinco y cuatro opciones de respuesta, lo que se traduce en un importante número de reactivos que no discriminan adecuadamente con cuatro opciones que con cinco.

Se considera que cuatro opciones de respuesta es lo más adecuado,[24-30] pero al observar las modificaciones en la discriminación y dificultad en cada reactivo, aunado a la mayor posibilidad de acierto al azar, consideramos que cinco opciones es una buena alternativa para la evaluación de altas consecuencias. Así que, el tiempo y esfuerzo invertidos en elaborar cinco opciones de respuesta en cada pregunta son recursos bien invertidos, siempre y cuando los elaboradores de las preguntas sean expertos en su área de conocimiento y tengan el apoyo de expertos en evaluación.

Asimismo, los índices dificultad, discriminación y comportamiento de los distractores son los más importantes para determinar la calidad de la prueba, además de ser las herramientas más útiles para los elaboradores al momento de calibrar el instrumento, al permitirles identificar errores o inconsistencias en la construcción y así corregirlas y generar una prueba con mayor

calidad y convertirla en una herramienta imprescindible en la retroalimentación de los cursos de especialización, en cuanto a sus fortalezas y debilidades.

REFERENCIAS

1. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-234.
2. Bruno JE, Dirkzwager A. Determining the optimal number of alternatives to a multiple-choice test item: An information Theoretic Perspective. Educational and Psychological Measurement 1995;55:959-966.
3. Haladyna TM, Downing, SM. How many options is enough for a multiple-choice test item? Educational and Psychological Measurement 1993;5:999-1010.
4. Haladyna TM, Downing, SM. A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 1989;2:37-50.
5. Ebel R. Expected reliability as a function of choices per item. Educational and Psychological Measurement 1969;29:565-570.
6. Ruch GM, Charles JW. A comparison of five types of objective tests in elementary psychology. J Applied Psychol 1928;12:398-404.
7. Muñiz J. Las teorías de los test: Teoría clásica y teoría de respuesta a los ítems. Papeles del Psicólogo 2010;31:57-66.
8. Costin F. The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. Educational and Psychological Measurement 1970;30:353-358.
9. Straton RG, Catts RM. A comparison of two, three and four choice item test given a fixed total number of choices. Educational and Psychological Measurement 1980;40:357-365.
10. Tversky A. On the optimal number of alternatives at a choice point. J Mathematical Psychology 1964;1:386-391.
11. Costin F. Three-choice versus four-choice items: implications for reliability and validity of objective achievement test. Educational and Psychological Measurement 1972;32:1035-1038.
12. Garner WR. Rating scales, discriminability and information transmission. Psychological Review 1960;67:343-352.
13. Grier JB. The number of alternatives for optimum test reliability. J Educational Measurement 1975;12:109-113.
14. Grier JB. The optimal number of alternatives at a choice point with travel time considered. J Mathematical Psychology 1976;12:31-97.
15. Rogers WT, Harley D. An empirical comparison of three- and four-choice items and test: Susceptibility to testwiseness and internal consistency reliability. Educational and Psychological Measurement 1999;59:234-247.
16. Olea J, Ponsoda V, Revuelta J, Hontangas P, Abad F. Analysis of the optimum number alternatives from the item response theory. Psicothema 2001;13:152-158.
17. Lord FM. Optimal number of choices per item: A comparison of four approaches. J Educational Measurement 1977;14:33-48.
18. Lissitz RW, Green SB. Effect of the number of scale points on reliability: A Monte Carlo approach. J Applied Psychology 1975;60:10-13.
19. Komorita SS, Graham WK. Number of scale points and the reliability of scales. Educational and Psychological Measurement 1965;25:987-995.
20. Masters JR. The relationship between number of response categories and reliability of Likert type questionnaires. J Educational Measurement 1974;11:49-53.
21. Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. Educational and Psychological Measurement 1971;31:657-674.
22. Peabody D. Two components in bipolar scales: Direction and extremeness. Psychological Review 1962;69:65-73.
23. Sancerni MD, Meliá JL y González V. Formato de respuesta, fiabilidad y validez, en la medición del conflicto de rol. Psicológica 1990;11:167-175.
24. Churchill GA Jr., Peter JP. Research design effects on the reliability of rating scales: A meta-analysis. J Marketing Research 1984;21:360-375.
25. Ferrando PJ. Saturaciones factoriales e índices de discriminación en la teoría clásica del test y en la teoría de respuesta al ítems. Anuario de Psicología 1994;2:55-65.
26. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica 2000;104:1-15.
27. Ramsay JO. The effect of number of categories in rating scales on precision of estimation of scale values. Psychometrika 1973;38:513-533.
28. Symonds PM. On the loss of reliability in ratings due to coarseness of the scale. J Experimental Psychology 1924;December:456-461.
29. Champney H, Marshall H. Optimal refinement of the rating scale. J Applied Psychology 1980;23:323-331.
30. Cox EP. The optimal number of response alternatives for a scale: A review. J Marketing Research 1980;17:407-422.
31. Aiken LR. Number of response categories and statistics on a teacher rating scale. Educational and Psychological Measurement 1983;43:397-401.
32. Bandalos DL, Enders CK. The effects of non-normality and number of response categories on reliability. Applied Measurement in Education 1996;9:151-160.
33. Cicchetti DV, Showalter D, Tyrer PJ. The effect of number of rating scale categories on levels of inter-rater reliability: A

Monte-Carlo investigation. Applied Psychological Measurement 1985;9:31-36.

34. Jenkins GD, Taber TD. A Monte Carlo study of factors affecting three indices of composite scale reliability. J Applied Psychology 1977;62:392-398.
35. Weng LJ. Impact of the number of response categories and anchor labels on coefficient alpha and test retest reliability. Educational and Psychological Measurement 2004;64:956-972.
36. Comrey AL, Montag I. Comparison of factor analytic results with two choice and seven choice personality item formats. Applied Psychological Measurement 1982;6:285-289.
37. Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika 1979;44:443-460.
38. Chang L. A psychometric evaluation of four-point and six-point Liker type scales in relation to reliability and validity. Applied Psychological Measurement 1994;18:205-215.
39. McCallum DM, Keith BR, Wiebe DJ. Comparison of response formats for multidimensional health locus of control scales: Six levels versus two levels. J Personality Assessment 1988;52:732-736.
40. Spearman C. General intelligence, objectively determined and measured. Am J Psychol 1904;15:202-293.
41. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15:72-101.
42. Spearman C. Demonstration of formulae for true measurement of correlation. Am J Psychol 1907;18:161-169.
43. Spearman C. Correlation calculated with faulty data. Br J Psychol 1910;3:271-295.
44. Spearman C. Correlations of sums and differences. Br J Psychol 1913;5:417-426.
45. Allen MJ, Yen WM. Introduction to measurement theory. Monterey, CA: Brooks/Cole, 1979,210-12.