Educación Médica 26 (2025) 101025

Educación Médica

www.elsevier.es/edumed

ORIGINAL

Comparación de indicadores psicométricos en evaluaciones de opción múltiple entre estudiantes de Medicina Humana

Adriana Villa Villavicencio*, Mariana Gómez Zamalloa, Diana Gabriela Ocsas Pinedo y César Gutiérrez

Facultad de Medicina Humana, Universidad de Piura, Lima, Perú

Recibido el 13 de noviembre de 2024; aceptado el 20 de enero de 2025
Disponible en Internet el 26 de febrero de 2025

PALABRAS CLAVE
Educación médica;
Preguntas de examen;
Psicometría;
Pedagogía;
Estudiantes de Medicina

Resumen
Introducción: las pruebas de opción múltiple son esenciales en la evaluación educativa, especialmente en medicina. Algunos estudios muestran que preguntas con 3, 4 o 5 alternativas tienen índices de dificultad y discriminación comparables.
Métodos: realizamos un estudio transversal con estudiantes del curso Metodología de la Investigación Científica II, quienes rindieron evaluaciones con 4 y 5 alternativas. Se calcularon los índices de dificultad y discriminación usando Excel y Jamovi. Para el análisis, se emplearon las pruebas de Kruskal-Wallis y coeficiente de correlación de concordancia. La prueba de chi-cuadrado fue utilizada para comparar distribuciones.
Resultados: participaron 55 alumnos. El índice de dificultad mostró medianas similares tanto para 4 como 5 alternativas (p = 0,824), al igual que el índice de discriminación que tampoco identificó diferencias entre las medianas (p = 0,654). El índice de dificultad tuvo buena concordancia en el examen parcial y final, con valores de 0,925 (IC 95%: 0,866 a 0,959) y 0,889 (IC 95%: 0,822 a 0,932), respectivamente. La concordancia para el índice de discriminación fue baja, con valores de 0,318 (IC 95%: −0,009 a 0,585) y −0,006 (IC 95%: −0,247 a 0,259) en el examen parcial y final, respectivamente.
Conclusión: no encontramos diferencias significativas en la dificultad y discriminación de preguntas de 4 y 5 alternativas. Sugerimos la elaboración de exámenes objetivos de respuesta única en cursos de investigación utilizando preguntas con 4 alternativas.

* Autor para correspondencia.

https://doi.org/10.1016/j.edumed.2025.101025

KEYWORDS
Medical education;
Examination questions;
Psychometrics;
Pedagogy;
Medical students

Comparison of psychometric indicators in multiple-choice evaluations with four and five options among medical students

Abstract
Introduction: Multiple-choice tests are essential in educational assessment, especially in medicine. Peruvian studies show that 3 and 5 alternatives have similar effectiveness, but the optimal number of alternatives remains debated globally.
Methods: We conducted a cross-sectional study with students from the assignment *"Metodología de la Investigación Científica II"*, who took assessments with four and five alternatives. Difficulty and discrimination index were calculated using Excel and Jamovi. Kruskal-Wallis test and concordance correlation coefficient where used fot he analysis, and chi-square test was used to compare distributions.
Results: A total of 55 students were included. The difficulty index showed similar medians for both 4 and 5 alternatives (p = 0.824), as did the discrimination index, which also did not identify differences between the medians (p = 0.654). The concordance values for the difficulty index indicated good consistency in both the midterm and final exams, with values of 0.925 (95% CI: 0.866 to 0.959) and 0.889 (95% CI: 0.822 to 0.932), respectively. The concordance for the discrimination index was low, with values of 0.318 (95% CI: −0.009 to 0.585) and −0.006 (95% CI: −0.247 to 0.259) for the midterm and final exams, respectively.
Conclusion: We did not find significant differences in the difficulty and discrimination of questions with 4 and 5 options. We suggest the use of 4 alternative questions in objective exams in research courses.

Introducción

A nivel mundial, las pruebas de opción múltiple se han convertido en una herramienta útil y sencilla para la evaluación de conocimientos[1]. En educación médica, su uso ha aumentado considerablemente, lo que ha impulsado investigaciones sobre su efectividad[1,2]. Debido a que es un método frecuente para medir el aprendizaje de los estudiantes, diferentes estudios han buscado determinar el número óptimo de alternativas por pregunta[2–4], utilizando indicadores como el índice de discriminación y de dificultad.

Actualmente, existe un extenso uso de preguntas con 5 alternativas, bajo la premisa de que un mayor número de alternativas aumenta la dificultad y refleja mejor la toma de decisiones en situaciones reales[2]. Sin embargo, algunos estudios sugieren que reducir el número de alternativas no afecta ni la dificultad ni la capacidad discriminativa del examen, además, ahorra tiempo a los estudiantes y examinadores en la realización y corrección de las pruebas, sin comprometer la calidad de la evaluación[2–5].

Identificamos varios estudios que comparan exámenes con preguntas de opción múltiple de 5, 4 y 3 alternativas, que fueron realizados en las asignaturas en ciencias básicas, clínicas y quirúrgicas, aplicados a estudiantes de pregrado y posgrado en carreras de salud[6–9]. No obstante, las variaciones en los contextos y asignaturas evaluadas en estos estudios dificultan la comparación de los resultados, puesto que los hallazgos en cursos de pregrado no necesariamente son aplicables al nivel de posgrado; ni los hallazgos en cursos de ciencias básicas son aplicables a cursos de clínica. En Perú, en una asignatura de bioquímica médica se encontró que exámenes escritos con preguntas de 3 y 5 alternativas mostraron similitudes en los índices de discriminación y dificultad[10].

No encontramos estudios que comparen exámenes escritos de 4 frente a 5 alternativas en asignaturas que enseñen métodos de investigación científica en la formación médica. Esta investigación tiene como objetivo comparar el índice de dificultad y discriminación de estos 2 tipos de preguntas en un grupo de estudiantes de Medicina Humana en un curso de Metodología de la Investigación de una universidad privada en Lima, Perú.

Material y métodos

Realizamos un estudio transversal y analítico en la facultad de medicina de una universidad privada en Lima, durante el primer semestre de 2024, en el contexto de la asignatura Metodología de la Investigación Científica II, obligatoria para los estudiantes de tercer año. La asignatura tiene una duración de 16 semanas e incluye 2 evaluaciones teóricas. El examen parcial se aplicó en la octava semana y el final al concluir las clases teóricas. Ambos exámenes evalúan conocimientos sobre el método científico en el diseño de proyectos de investigación, ética en investigación, análisis crítico de la literatura y estrategias metodológicas para controlar sesgos y efectos confusores.

Diseño muestral

Decidimos no calcular un tamaño de muestra, ya que trabajamos con la población total de alumnos matriculados en Metodología de la Investigación Científica II, estos fueron

56. El criterio de inclusión fue estar matriculado en dicha asignatura, excluimos aquellos estudiantes que participaron únicamente en una de las 2 evaluaciones. El análisis final incluyó a 55 alumnos.

Definición de variables

La principal unidad de análisis fueron las preguntas de los exámenes parcial y final, que sumaron un total de 100. Estos incluían 2 tipos de preguntas de opción múltiple, con 4 y 5 alternativas, distribuidas en igual proporción. Analizamos cada pregunta mediante el índice de dificultad y el índice de discriminación.

El índice de dificultad cuantifica la proporción de respuestas correctas entre el total de respuestas para cada pregunta[11], este permite evaluar el número de alumnos que seleccionaron la respuesta correcta sobre el total de alumnos que participaron en la resolución de la pregunta[1]. Utilizamos los siguientes puntos de corte para determinar el grado de dificultad: mayor o igual que 85% como fácil; entre 51 y 84% como moderado y menor que 50%, difícil[12].

Por otro lado, el índice de discriminación mide la diferencia entre la proporción de alumnos de alto rendimiento (aquellos que obtuvieron puntajes elevados en el examen) que respondieron correctamente la pregunta frente a los alumnos de bajo rendimiento que también lo hicieron[11]; este índice fue estimado con el coeficiente de correlación biserial puntual[13]. Los puntos de corte fueron: menor que 0,19 para baja discriminación, entre 0,20 y 0,29 insuficiente, entre 0,30 y 0,39 buena, y más de 0,40 muy buena[14].

Procedimientos del estudio

Construcción de los exámenes

La elaboración de las preguntas de opción múltiple con 4 y 5 alternativas fue realizada por el jefe de la asignatura de Metodología de la Investigación Científica II. Todas las preguntas originalmente contenían 5 alternativas, para su reducción a 4 alternativas se eliminó un distractor de manera aleatoria. Las preguntas se presentaron en los exámenes parcial y final, donde cada uno tuvo 2 tipos de examen: A y B (fig. 1). Aunque este fue el planteamiento inicial, en la versión final de los exámenes se distribuyeron aleatoriamente las preguntas entre ambos tipos de examen.

Aplicación de la prueba

Ambas pruebas fueron aplicadas como parte de las evaluaciones programadas en el sílabo, que establece una nota mínima aprobatoria de 11 en la escala vigesimal. Los alumnos conformaron una sola aula y rindieron las evaluaciones el mismo día y en el mismo horario, utilizando un cuadernillo y ficha óptica. Fueron ubicados en columnas según el orden de la lista y se les asignó uno de los 2 tipos de examen según su ubicación. Este proceso fue aplicado para ambos exámenes.

Recolección de datos

Solicitamos a la oficina de evaluaciones la base de datos con las respuestas correctas y las alternativas seleccionadas por los alumnos en los exámenes parcial y final; esta se encargó de la corrección, asignación de calificaciones y seudoanonimización de los datos. Organizamos las preguntas en columnas en una hoja de Excel. Registramos los aciertos de cada pregunta junto al código de identificación del alumno, aplicamos este proceso en ambos exámenes.

Para comparar los puntajes obtenidos por cada alumno en preguntas con 4 y 5 alternativas, creamos una base de datos separada en Excel. Registramos los aciertos de cada alumno, dividiéndolos en aciertos totales en preguntas de 4 y 5 alternativas dentro de la misma evaluación, ya fuera parcial o final.

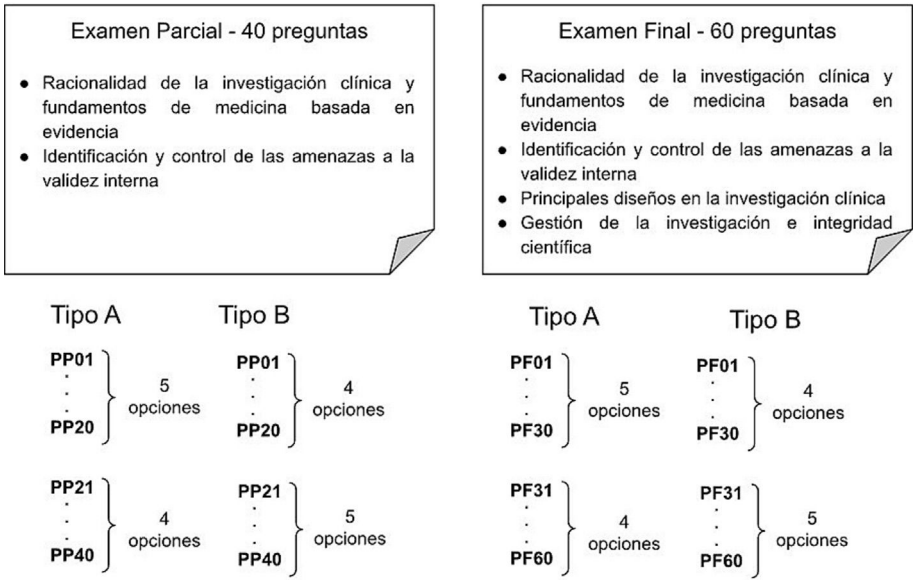

Figura 1 Procedimiento de elaboración del examen parcial y final.

Análisis de datos

Utilizamos Excel para calcular el índice de dificultad de cada pregunta 2 veces: primero para la versión de 4 alternativas y luego para la de 5. El índice de discriminación fue calculado con el programa Jamovi en el módulo snowRMN (versión 2.3.21). Analizamos los datos por separado en los exámenes parcial y final, y luego según el número de alternativas, generando 4 grupos de comparación.

Reingresamos los valores obtenidos para ambos índices en Jamovi y aplicamos la prueba Kruskal-Wallis a los 4 grupos para analizar las medianas. Calculamos el coeficiente de correlación de concordancia, disponible en el módulo SimplyAgree de Jamovi, similar al coeficiente de correlación intraclase[15,16], este evalúa la concordancia entre mediciones cuantitativas, con valores entre 0 y 1: 0 indica mayor variabilidad y 1 indica máxima concordancia. Consideramos la concordancia muy buena si el valor fue mayor que 0,90; buena entre 0,71 y 0,90; moderada entre 0,51 y 0,70; mediocre entre 0,31 y 0,50 y mala o nula si fue inferior a 0,31[17].

Clasificamos los índices de dificultad y discriminación para cada pregunta según las categorías definidas en la sección de variables. Usamos la prueba de chi-cuadrado para comparar la distribución de estas categorías entre los 4 grupos. Además, en la base de datos de aciertos por alumno, calculamos el promedio de aciertos individuales por tipo de pregunta y comparamos las medianas para cada número de alternativas con la prueba de Wilcoxon. Luego, obtuvimos el coeficiente de Spearman para evaluar la correlación entre la proporción de aciertos de cada alumno para cada uno de los tipos de exámenes.

Finalmente, plasmamos la concordancia entre las respuestas de las preguntas con 4 y 5 alternativas en gráficos de Bland–Altman, útiles para evaluar la concordancia entre 2 técnicas de medición y determinar si presentan coincidencia suficiente como para considerarse intercambiables[18].

Para todas las pruebas de hipótesis estadísticas aplicadas, consideramos un nivel de significación de 0,05.

Resultados

En el examen parcial, 37 de 55 alumnos (67,27%) aprobaron, mientras que en el examen final lo hicieron 40 de 55 (72,73%). Las calificaciones del parcial variaron entre 6,5 y 14,5, con una mediana de 11,5; en contraste, en el examen final, las notas oscilaron entre 3,0 y 16,3, con una mediana de 11,6.

Índice de dificultad

Para calcular el índice de dificultad incluimos las 100 preguntas y determinamos las medianas y los rangos intercuartílicos (RI) para los 4 grupos. En el examen parcial, la mediana de aciertos fue 0,565 (RI = 0,315 a 0,829) para preguntas de 4 alternativas y 0,542 (RI = 0,321 a 0,757) para las de 5. En el examen final, la mediana fue 0,648 (RI = 0,404 a 0,780) para preguntas de 4 alternativas y 0,573 (RI = 0,423 a 0,786) para las de 5 (fig. 2A). Aplicamos la prueba de Kruskal-Wallis para evaluar diferencias significativas entre las medianas de los índices en preguntas de 4 y 5 alternativas en ambos exámenes, sin encontrar diferencias entre los grupos (p = 0,824).

En cuanto a la distribución de las preguntas según el grado de dificultad, en el examen parcial el 25% de las preguntas de 4 alternativas fueron clasificadas como fáciles, el 30% moderadas y el 45% difíciles. Para las preguntas de 5 alternativas en el parcial, el 13% fueron fáciles, el 45% moderadas y el 43% difíciles. En el final, el 13% de las preguntas de 4 alternativas fueron fáciles, el 54% moderadas, y el 33% difíciles; mientras que para las de 5 alternativas, el 13% fueron fáciles, el 48% moderadas y el 38% difíciles (fig. 2B). Además, no encontramos asociación entre la dificultad y el tipo de examen (p = 0,326).

Calculamos el coeficiente de correlación de concordancia entre los índices de dificultad de ambos modelos de pregunta para los exámenes parcial y final de manera individual, obteniendo una concordancia «muy buena» en el parcial (fig. 2C) y «buena» en el final (fig. 2D).

Índice de discriminación

Analizamos 94 preguntas en total, tras eliminar 6 que tuvieron un valor constante, es decir, todos los alumnos marcaron la alternativa correcta. En el examen parcial, la mediana fue 0,273 (RI = 0,150 a 0,454), para preguntas de 4 alternativas y 0,332 (RI = 0,163 a 0,423) para las de 5. En el examen final, la mediana fue 0,362 (RI = 0,187 a 0,440) para preguntas de 4 alternativas y 0,324 (RI = 0,219 a 0,482) para las de 5 (fig. 3A**)**. Al comparar las medianas con las pruebas Kruskal-Wallis, no encontramos una diferencia entre los grupos (p = 0,654).

En el examen parcial, el 31% de las preguntas de 4 alternativas y el 31% de las de 5 alternativas mostraron una discriminación muy buena. En las preguntas de 4 alternativas, el 14% fue buena, el 23% insuficiente y el 31% mala; mientras que en las de 5 alternativas, el 23% fue buena, el 17% insuficiente y el 29% mala. En el examen final, el 46% de las preguntas de 4 alternativas mostraron discriminación muy buena, el 14% buena, el 14% insuficiente y el 27% mala; para las preguntas de 5 alternativas, el 36% tuvieron discriminación muy buena, el 24% buena, el 19% insuficiente y el 22% mala (fig. 3B). Al aplicar la prueba chi-cuadrado, no encontramos diferencias significativas entre las distribuciones (p = 0,727).

Finalmente, calculamos el coeficiente de correlación de concordancia entre los índices de ambos modelos de pregunta. Obtuvimos una concordancia «mediocre» en el parcial y «nula» en el final, con coeficientes de 0,318 (IC 95%: −0,009 a 0,585) (fig. 3C) y −0,006 (IC 95%: −0,247 a 0,259) (fig. 3D), respectivamente.

Diferencia entre tipos de examen en el mismo alumno

Al analizar el rendimiento de cada alumno cuando responde preguntas con 4 o con 5 alternativas, la mediana de aciertos en las preguntas de 4 alternativas fue 0,600 y en las de 5 fue 0,550. Al realizar la prueba de Wilcoxon, encontramos una diferencia significativa entre las medianas de los grupos (p < 0,05). La correlación en el examen parcial fue

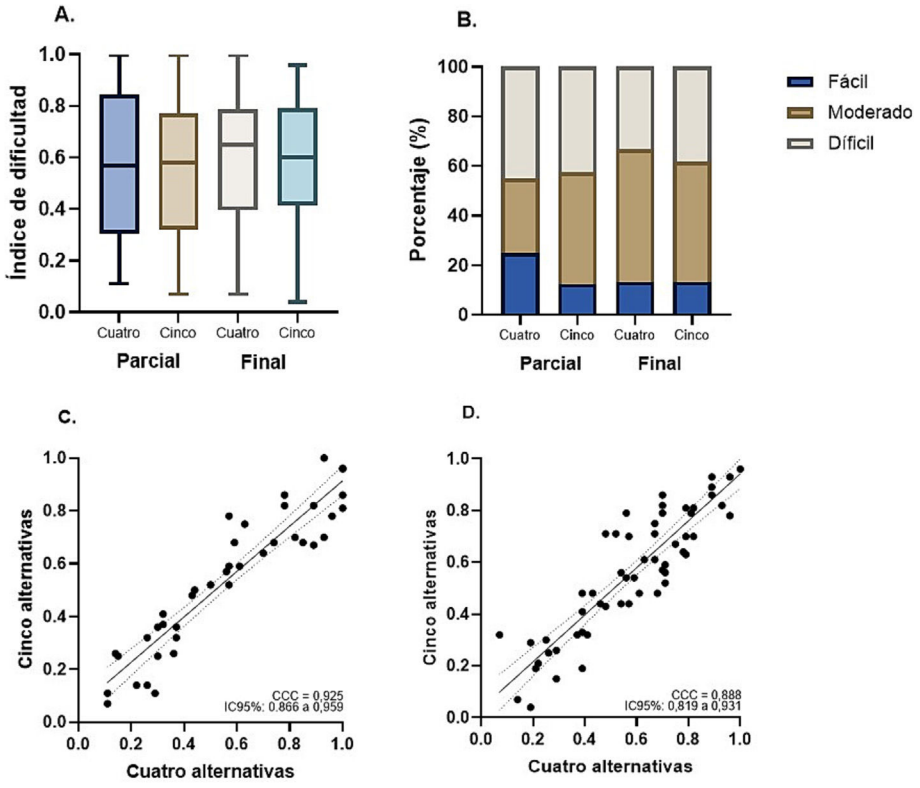

Figura 2 **A)** Comparación de índices de dificultad de las preguntas de 4 versus 5 alternativas en las pruebas parcial y final. **B)** Distribución de las preguntas según nivel de dificultad. **C)** Gráfico de dispersión de los índices de dificultad de las preguntas de 4 y 5 alternativas en el examen parcial. **D)** Gráfico de dispersión de los índices de dificultad de las preguntas de 4 y alternativas en el examen final. CCC: Coeficiente de correlación de concordancia.

moderada positiva, con un coeficiente de 0,418 (IC 95%: 0,164 a 0,260) (fig. 4A), mientras que en el examen final la correlación fue fuerte con un coeficiente de 0,651 (IC 95%: 0,459 a 0,784) (fig. 4C).

En los gráficos de Bland–Altman se observa que en el examen parcial hubo mayor variación en las diferencias entre las respuestas de 4 y 5 alternativas, lo que se visualiza como una mayor dispersión (fig. 4B). En contraste, el examen final muestra que, si bien hubo variabilidad, las diferencias entre ambos tipos de pregunta son más simétricas y evidencian mejor concordancia (fig. 4D).

Discusión

Nuestros hallazgos mostraron medianas similares para los índices de dificultad y discriminación en los 4 grupos, sugiriendo que no hay diferencia entre pruebas de 4 y 5 alternativas en la evaluación de conocimientos en metodología de investigación científica. Esto sugiere que los exámenes podrían optimizarse utilizando solo 4 alternativas, permitiendo reducir el número de distractores sin afectar la calidad de la evaluación.

Al analizar el grado de dificultad de las preguntas, observamos que la mayoría se clasificaron entre niveles «moderado» y «difícil» en ambos exámenes, mostrando una distribución homogénea en las 4 pruebas y sin diferencias en las medianas de los índices de dificultad. Estos resultados coinciden con otros estudios en estudiantes del área de la salud, que también encontraron que reducir las alternativas de 5 a 4 o a 3 no afecta el índice de dificultad en las evaluaciones[1,7,8].

En conjunto, la distribución de la dificultad de las preguntas se alejó ligeramente de los niveles óptimos sugeridos[19]. Las preguntas fáciles representaron entre un 12,5 y 25%, mientras que lo recomendado es un 5%. Las preguntas de dificultad moderada oscilaron entre el 30 y 53,3%, acercándose al valor óptimo del 70%. Sin embargo, las preguntas difíciles oscilaron entre un 33,3 y 45%, superando el 25% sugerido, lo que indica una mayor proporción de preguntas desafiantes.

No encontramos diferencias entre las medianas del índice de dificultad de ambas pruebas, esto concuerda con un estudio en estudiantes de primer año de Medicina en las asignaturas de Química, Anatomía y Fisiología, en dichos exámenes no hubo diferencias en el índice de discriminación entre preguntas de 4 y 5 alternativas[7]. Un estudio de 2018 reportó una diferencia mínima en los índices de discriminación entre preguntas de 3, 4 y 5 alternativas en la asignatura de fisiología, mostrando índices de 0,25 para las preguntas de 3 y 4 alternativas y 0,27 para las de 5[6]. Los autores atribuyen esta diferencia a un hallazgo incidental relacionado con preguntas consideradas fáciles.

Entre 45,7 y 54,3% de las preguntas en el examen parcial y entre 59,3 y 59,4% en el examen final tuvieron buena

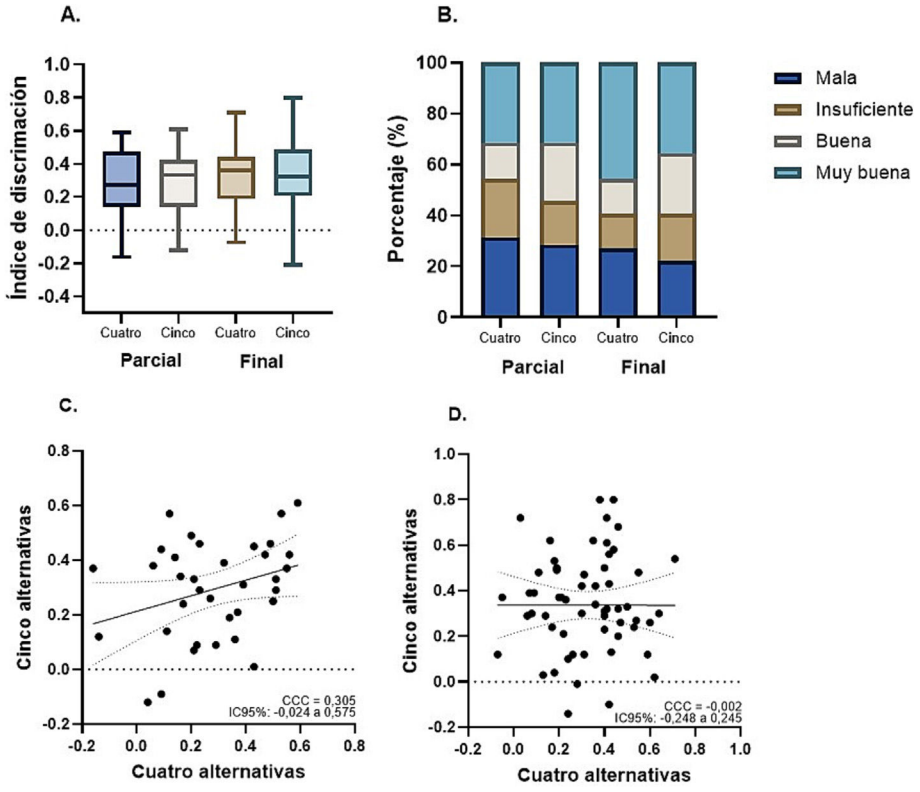

Figura 3 **A)** Comparación de índices de discriminación de las preguntas de 4 versus 5 alternativas en las pruebas parcial y final. **B)** Distribución de las preguntas según nivel de discriminación. **C)** Gráfico de dispersión de los índices de discriminación de las preguntas de 4 y 5 alternativas en el examen parcial. **D)** Gráfico de dispersión de los índices de discriminación de las preguntas de 4 y alternativas en el examen final. CCC: Coeficiente de correlación de concordancia.

discriminación. A diferencia de un estudio en alumnos de último año de Medicina donde se reportó 77,8% de preguntas con discriminación «pobre» y «negativa» en preguntas de 4 y 5 alternativas, posiblemente debido a la calidad de los distractores[8]. Es fundamental analizar los factores externos como la preparación de los alumnos, la asignatura evaluada, entre otros, para lograr una discriminación efectiva en todas las preguntas. Un índice entre 0,30 y 0,40 es aceptable, sin embargo, la pregunta podría ser susceptible de mejorar; un índice superior a 0,40 se considera excelente[20].

Con respecto al índice de dificultad, este mostró un alto coeficiente de correlación de concordancia en ambos exámenes, indicando que ambos formatos de pregunta coinciden en nivel de dificultad. Sin embargo, el índice de discriminación presentó baja concordancia, lo que indica que las preguntas no discriminan uniformemente en ambas evaluaciones. Esto sugiere que las preguntas no lograron diferenciar de manera consistente entre los estudiantes con distintos niveles de rendimiento. Cabe resaltar que el índice de discriminación considera factores externos, como el rendimiento general del alumno, mientras que el índice de dificultad solo mide el porcentaje de respuestas correctas en cada pregunta.

El rendimiento de los alumnos frente a ambos números de alternativas en un mismo examen mostró una correlación moderada positiva en el parcial y fuerte positiva en el final. Los gráficos de Bland–Altman indican que las diferencias en puntajes disminuyeron en el examen final. Hipotetizamos que esto podría deberse a que el examen parcial cubrió temas que se evaluaron nuevamente en el final, permitiendo afianzar conocimientos y mejorar la correlación en el último examen.

Nuestra principal limitación es el tamaño de la muestra, al incluir solo 55 alumnos. Además, el estudio se realizó en una única institución educativa, en un solo curso de métodos de investigación científica aplicada a la medicina dentro de un año académico específico. Esto contrasta con otros estudios que evaluaron la efectividad del número de alternativas en diversos cursos[7,8], por lo que se reduce la posibilidad de generalizar nuestros resultados. Por otro lado, optamos no realizar el análisis de la funcionalidad de los distractores ya que requiere un estudio específico.

Una fortaleza fue la inclusión equitativa de preguntas de opción múltiple con 4 y 5 alternativas en el examen parcial y final, lo que minimizó sesgos en la percepción de dificultad. Además, realizamos un análisis individual de los índices de dificultad y discriminación de cada pregunta y usamos a cada alumno como su propio control para comparar su rendimiento.

Los resultados de este estudio de comparación y análisis de los índices de dificultad y discriminación respaldan que la reducción del número de alternativas de 5 a 4 en las preguntas de opción múltiple puede ser una estrategia para optimizar la elaboración de exámenes objetivos de

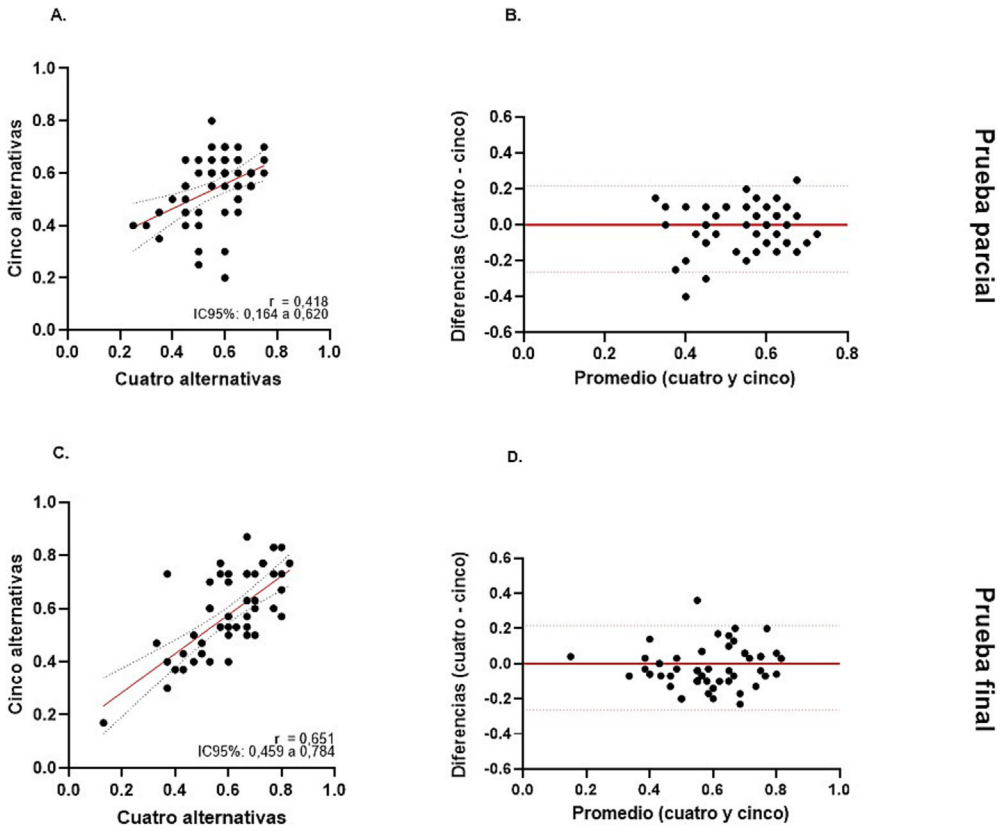

Figura 4 **A)** Análisis de correlación entre la proporción de respuestas correctas entre las preguntas de 4 versus 5 alternativas respondidas por un mismo estudiante en el examen parcial. **B)** Gráfico de Bland–Altman para representar la diferencia de la proporción de preguntas de 4 versus 5 alternativas correctamente respondidas por un mismo estudiante en el examen parcial. **C)** Análisis de correlación entre la proporción de respuestas correctas entre las preguntas de 4 versus 5 alternativas respondidas por un mismo estudiante en el examen final. **D)** Gráfico de Bland–Altman para representar la diferencia de la proporción de preguntas de 4 versus 5 alternativas correctamente respondidas por un mismo estudiante en el examen final.

respuesta única en cursos de investigación en estudiantes de Medicina.

Responsabilidades éticas

Comité Institucional de Ética en Investigación de la Universidad de Piura (expediente N° T0324–05).

Consentimiento informado

No se hizo uso del consentimiento informado ya que las evaluaciones que se utilizaron para obtener los datos no fueron aplicadas directamente por nosotros, los investigadores, estas formaron parte del curso de Metodología de la Investigación Científica II. Una vez aprobado nuestro protocolo por el comité de ética, la secretaría académica de la institución nos brindó una base de datos seudoanonimizada que contuvo el desarrollo del examen y las calificaciones de los participantes. Se informó a los estudiantes del curso de Metodología de la Investigación II sobre la estructura que tuvieron los exámenes que se rindieron a través del sílabo del curso y que dicha información sería utilizada para este estudio.

Financiación

Los autores declaran haber recibido financiación de parte de la Facultad de Medicina Humana de la Universidad de Piura.

Conflicto de intereses

Los autores han recibido ayuda de la Facultad de Medicina Humana de la Universidad de Piura, Lima, Perú.

Agradecimientos

A Franco Romaní de la Unidad de Investigación de la Facultad de Medicina por su apoyo en la revisión del manuscrito y en la elaboración de los gráficos.

Bibliografía

1. Al-lawama M, Kumwenda B. Decreasing the options' number in multiple choice questions in the assessment of senior medical students and its effect on exam psychometrics and distractors' function. BMC Med Educ. 2023;23(1):212.

2. Esmaeeli B, Esmaeili Shandiz E, Norooziasl S, Shojaei H, Pasandideh A, Khoshkholgh R, et al. The optimal number of choices in multiple-choice tests: a systematic review. Med Educ Bull. 2021;2(3):253–60.
3. Schneid SD, Armour C, Park YS, Yudkowsky R, Bordage G. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. Med Educ. 2014;48(10):1020–7.
4. Rahma NAA, Shamad MMA, Idris MEA, Elfaki OA, Elfakey WEM, Salih KMA. Comparison in the quality of distractors in three and four options type of multiple choice questions. Adv Med Educ Pract. 2017;8:287–91.
5. Vegada B, Shukla A, Khilnani A, Charan J, Desai C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: a randomized study. Indian J Pharmacol. 2016;48(5):571.
6. Loudon C, Macias-Muñoz A. Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: implications for exam design. Adv Physiol Educ. 2018;42(4):565–75.
7. Fozzard N, Pearson A, du Toit E, Naug H, Wen W, Peak IR. Analysis of MCQ and distractor use in a large first year Health Faculty Foundation Program: assessing the effects of changing from five to four options. BMC Med Educ. 2018;18(1):252.
8. Belay LM, Sendekie TY, Eyowas FA. Quality of multiple-choice questions in medical internship qualification examination determined by item response theory at Debre Tabor University, Ethiopia. BMC Med Educ. 2022;22(1):635.
9. Kheyami D, Jaradat A, Al-Shibani T, Ali FA. Item analysis of multiple choice questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. Sultan Qaboos Univ Med J. 2018;18(1):e68–74.
10. Camacho-Saavedra LA, Huamán-Saavedra JJ, Plasencia-Alvarez JO. Eficiencia de preguntas de opción múltiple con 3 alternativas. Rev Médica Trujillo [Internet] 2020 15(4) [consultado 1 Nov 2024]. Disponible en: https://revistas.unitru.edu.pe/index.php/RMT/article/view/3214.
11. Charles S, Denison DB. Handbook on measurement, assessment, and evaluation in higher education [Internet] 2.ª ed. 2017 [consultado 1 Nov 2024]. Disponible en: https://www.routledge.com/Handbook-on-Measurement-Assessment-and-Evaluation-in-Higher-Education/Secolsky-Denison/p/book/9781138892156.
12. Understanding item analyses | Office of Educational Assessment [Internet] [consultado 1 Nov 2024]. Disponible en: https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/ 2001.
13. Billings MS, DeRuchie K, Hussie K, GUÍA DE REDACCIÓN DE PREGUNTAS DEL NBME. Elaboración de preguntas para evaluaciones escritas en el área de la salud. [Internet] [consultado 1 nov 2024]. Disponible en: https://www.nbme.org/sites/default/files/2022-10/NBME_Item-Writing_Guide_Spanish.pdf 2022.
14. Elosua Oliden P, Egaña M. Psicometría aplicada. Guía para el análisis de datos y escalas con jamovi [Internet].Servicio Editorial de la Universidad del País Vasco/Euskal Herriko Unibertsitatearen Argitalpen Zerbitzua; 2020 [consultado 1 nov 2024]. Disponible en: http://addi.ehu.es/handle/10810/43054.
15. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255–68.
16. Caldwell AR. SimplyAgree: an R package and jamovi module for simplifying agreement and reliability analyses. J Open Source Softw. 2022;7(71):4148.
17. Argimon Pallás J. Métodos de investigación clínica y epidemiológica. 5ta ed. España: Elsevier; 2019;522.
18. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurements. The Lancet. 1986;327(8476):307–10.
19. Hurtado L. Dificultad asignada y observada de las preguntas de una prueba de rendimiento académico. Rev Estud Psicológicos. 2021;1(4):80–101.
20. Ebel R.L. and Frisbie D.A., Essentials of educational measurement [Internet], 5ta ed., 1991, 363, [Consultado 1 Nov 2024]. Disponible en: https://ebookppsunp.files.wordpress.com/2016/06/robert_l-ebel_david_a-_frisbie_essentials_of_edbookfi-org.pdf.