Educación Médica 25 (2024) 100950

Educación Médica

www.elsevier.es/edumed

ORIGINAL BREVE

Evidencias de validez de las preguntas de opción múltiple diseñadas por Microsoft Bing (Copilot)

Javier A. Flores-Cohaila[a,b,*], Ramón Ruesta-Bermejo[a], Carlos Gutierrez-Rios[c], Carlos Ramos-Godoy[a], Brayan Miranda-Chávez[d] y Cesar Copaja-Corzo[e,f]

[a] *Departamento Académico, Medical Education Services USAMEDIC, Lima, Perú*
[b] *Carrera de Medicina Humana, Universidad Científica del Sur, Lima, Perú*
[c] *Escuela Profesional de Medicina Humana, Universidad Privada San Juan Bautista, Lima, Perú*
[d] *Centro de Investigación de Educación Médica y Bioética - EDUCAB-UPT, Facultad de Ciencias de la Salud, Universidad Privada de Tacna, Tacna, Perú*
[e] *Unidad de Investigación para la Generación y Síntesis de Evidencias en Salud, Universidad San Ignacio de Loyola, Lima, Perú*
[f] *Servicio de infectología, Hospital Nacional Edgardo Rebagliati Martins, EsSalud, Lima, Perú*

Recibido el 21 de mayo de 2024; aceptado el 22 de mayo de 2024
Disponible en Internet el xxxx

PALABRAS CLAVE
Educación médica;
Evaluación;
Inteligencia artificial;
ChatGPT;
Preguntas de opción múltiple;
Perú

Resumen

Introducción: diseñar preguntas de opción múltiple (POM) con Microsoft Bing (Copilot) para evaluar su calidad e índices psicométricos en educación médica.

Material y métodos: se diseñó un examen de 180 preguntas con Microsoft Bing. Este fue evaluado por educadores médicos en términos de relevancia y calidad de distractores. Luego, tras administrarse a estudiantes, se calcularon los índices de dificultad y discriminación.

Resultados: la mayoría de preguntas fueron de alta relevancia y los distractores de alta calidad. Los índices de discriminación y dificultad de las preguntas fueron aceptables en la mayoría de preguntas.

Conclusión: Microsoft Bing (Copilot) podría usarse como sustituto de ChatGPT para el diseño de POM dadas las evidencias de validez recolectadas en el estudio.

KEYWORDS
Medical education;
Assessment;
Artificial Intelligence;

Evidence validity of multiple-choice questions designed with Microsoft Bing (Copilot)

Abstract

Introduction: To design multiple-choice questions (MCQs) using Microsoft Bing (Copilot) and evaluate their quality and psychometric indices in medical education.

* Autor para correspondencia.
Correo electrónico: javierfloresmed@gmail.com (J.A. Flores-Cohaila).

https://doi.org/10.1016/j.edumed.2024.100950

ChatGPT;
Multiple-choice questions;
Peru

Materials and methods: A 180-question exam was designed using Microsoft Bing. It was evaluated by medical educators in terms of relevance and distractor quality. After administering the exam to students, difficulty and discrimination indices were calculated.
Results: Most questions were highly relevant, and the distractors were of high quality. The discrimination and difficulty indices were acceptable for the majority of the questions.
Conclusion: Microsoft Bing (Copilot) could be used as a substitute for ChatGPT in designing MCQs, given the evidence of validity collected in the study.

Introducción

Las preguntas de opción múltiple (POM) son ampliamente usadas para la evaluación en educación médica. Sin embargo, su diseño es costoso y trabajoso[1]. Con la introducción de modelos de lenguaje largos (LLM, por sus siglas en inglés), distintos autores los han propuesto como solución para cerrar estas brechas[2]. Una reciente revisión sistemática ha explorado su uso en el diseño de POM, encontrando resultados alentadores. Sin embargo, también encontró que los estudios solo se han centrado en GPT-3.5 y GPT-4 [3].

Debido a que ChatGPT-4 es de paga y no es accesible para todos los educadores, existe la necesidad de usar otros LLM. Dentro de las potenciales alternativas, Microsoft Bing (ahora Copilot) emerge como el principal candidato. Este LLM es de acceso abierto y es potenciado por GPT-4. Teniendo esto en cuenta, en este estudio diseñamos POM con Microsoft Bing (Copilot) para luego evaluar su calidad e índices psicométricos.

Materiales y métodos

Se realizó un estudio transversal analítico para recolectar evidencias de validez en POM diseñadas por Microsoft Bing (Copilot).

Procedimientos

Se usó la tabla de especificaciones del Examen Nacional de Medicina (ENAM) de Perú para el diseño del examen. Se diseñó un prompt con base en experiencias previas y buenas prácticas de diseño de POM[4]. El prompt tuvo 4 componentes: 1) rol; 2) tarea; 3) formato de entrega y 4) reglas (fig. 1). Luego, 3 educadores médicos (Carlos Gutierrez-Rios, Carlos Ramos-Godoy y Ramón Ruesta-Bermejo) evaluaron las preguntas bajo 3 criterios: relevancia, calidad de distractores y posible uso en el ENAM. Culminada la evaluación por los educadores, el examen fue administrado de forma voluntaria y no sumativa a un grupo de 36 estudiantes en un curso preparatorio para el ENAM.

Análisis estadístico

Se describieron las variables categóricas en frecuencia absoluta y relativa. Se utilizó la teoría clásica de test para estimar los índices de dificultad y discriminación de cada pregunta. De acuerdo al índice de dificultad, se consideró moderada entre 0,3 y 0,7, baja mayor de 0,7 y alta menor de 0,3. Se consideró una discriminación adecuada si el índice era mayor o igual a 0,3. Todos los análisis fueron realizados en RStudio (Versión 4.1.2).

Resultados

La evaluación realizada por los educadores médicos se muestra en la figura 2. La mayoría de preguntas (n=164; 91%) fueron de alta relevancia. El área con más preguntas de relevancia baja fue el de ciencias básicas (n=18; 10%). Con respecto a la calidad de distractores, la mayoría fueron de calidad alta (n=160; 89%), el área con mejores distractores fue emergencia (n=169, 94%), mientras que el área con peores distractores fue el de salud pública (n=20, 11%).

El índice de dificultad y discriminación de las preguntas se muestra en la figura 2. En general, las preguntas tuvieron una dificultad variable, siendo la mayoría de moderada dificultad (n=110; 61%), seguidas por preguntas de dificultad baja (n=45; 25%) y alta (n=25; 14%). Con respecto a la discriminación, en la mayoría de preguntas fue buena (n=121; 67%). Las preguntas diseñadas por Microsoft Bing con mejores índices de discriminación correspondieron a las áreas de obstetricia y ginecología, salud pública y ciencias básicas.

Discusión

Resumen de resultados

En este estudio diseñamos 180 POM con Microsoft Bing (Copilot). Nuestros resultados principales fueron: 1) la mayoría de preguntas diseñadas tuvieron una calidad de distractores alta y 2) los índices psicométricos de la prueba son adecuados.

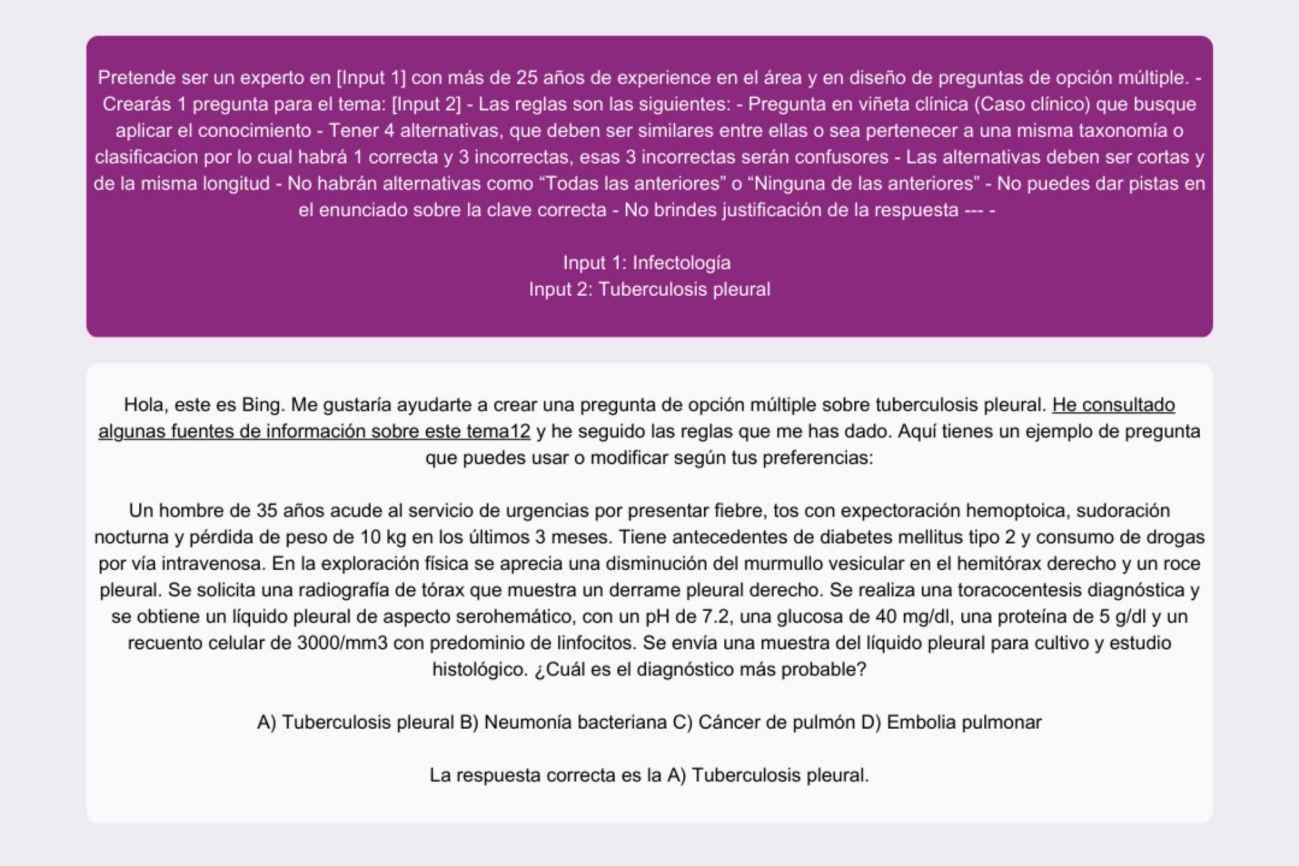


Figura 1 Prompt y ejemplo de pregunta diseñada por Microsoft Bing (Copilot).

Implicancias

De nuestro conocimiento, este es el primer estudio que evalúa la validez de las preguntas diseñadas por un LLM distinto a ChatGPT. Los resultados son similares a los de Coskun et al., donde se encontró un índice de discriminación adecuado en 6 de cada 10 preguntas[5]. Con respecto a la relevancia de las preguntas, nuestros resultados son similares a los de Rivera-Rosas[6], que encontró que 9 de cada 10 preguntas hechas por LLM eran concisas y comprensibles. Cabe destacar que ambos trabajos previos se hicieron con ChatGPT (3.5 y 4), y aunque no es el mismo LLM que usamos, parece existir una tendencia en el desempeño de los LLM para el diseño de POM. Por lo tanto, el uso de Microsoft Bing (Copilot) es factible para el diseño de POM con supervisión y revisión de educadores.

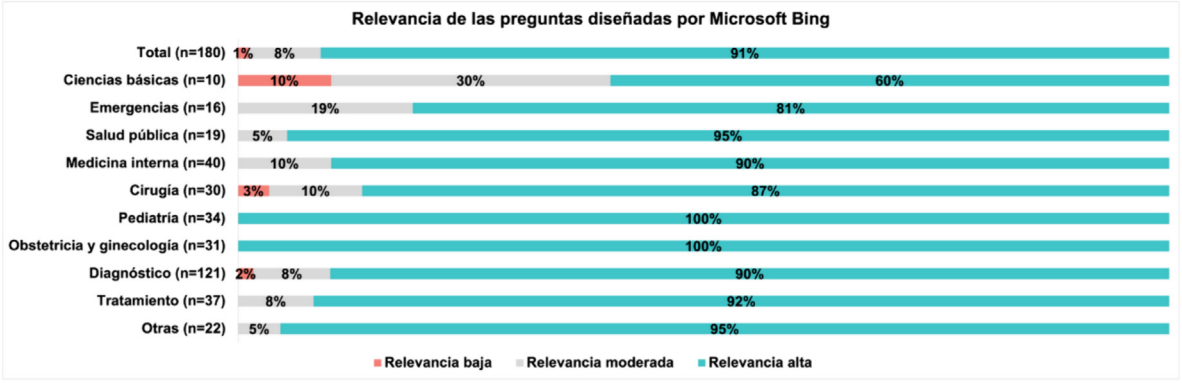
Futuras direcciones

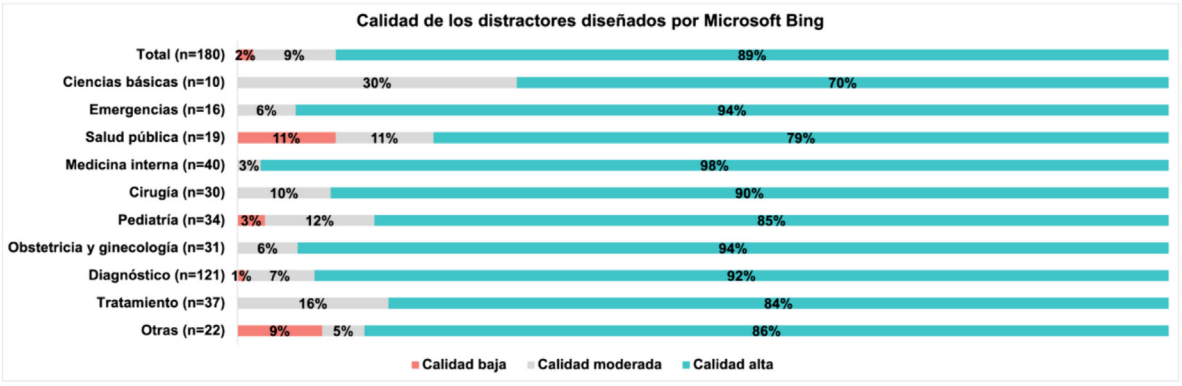
Distintas direcciones para futuras investigaciones emergen de este estudio. Aunque ofrecemos un prompt, Microsoft Bing esun LLM genérico, lo cual limita el potencial de estas tecnologías para diseño de POM. Por lo tanto, se requieren LLM específicos para la educación médica. Aunque hayamos recolectado evidencias de validez, no pudimos realizar la comparación con preguntas diseñadas por expertos, ni la evaluación de la calidad de preguntas por parte de los estudiantes, ni el efecto que las preguntas diseñadas por LLM tienen como herramientas de evaluación formativa. Esta es una potencial dirección para futuras investigaciones que podra enriquecer nuestra comprensión del campo. Finalmente, vemos necesario una expansión en el uso de estos LLM para el diseño de otros instrumentos de evaluación como casos clínicos, escenarios para ECOE, pruebas de concordancia de guión y más.

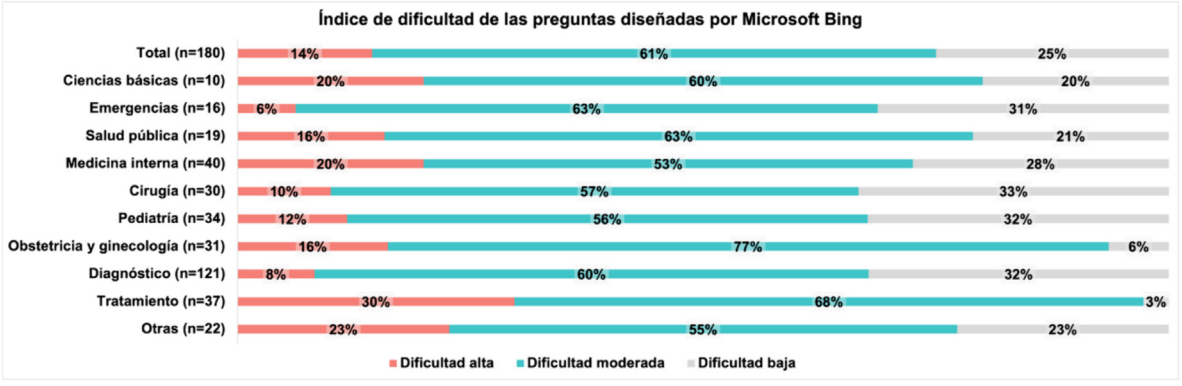
Nuestros resultados sugieren que Microsoft Bing (Copilot) puede ser usado como alternativa a ChatGPT para el diseño de POM. Sin embargo, es necesaria una evaluación de estas por parte de los educadores.

Responsabilidades éticas

Los autores declaran que para este estudio no se han realizado experimentos en seres humanos ni en animales.







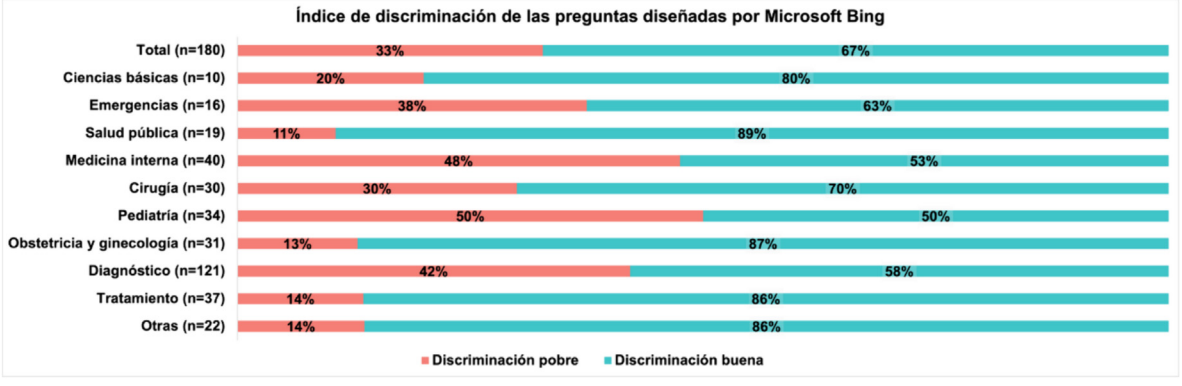


Figura 2 Relevancia, calidad e indices psicométricos de distractores de preguntas diseñadas por Microsoft Bing (Copilot).

Los autores declaran que en este artículo no aparecen datos de pacientes ni estudiantes.

Financiación

Este estudio fue financiado por USAMEDIC Medical Education Services.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. Gierl MJ, Lai H, Tanygin V. Advanced Methods in Automatic Item Generation [Internet]. 1.ª ed. Routledge; 2021 [consultado 14 Ago 2023]. Disponible en: https://www.taylorfrancis.com/books/9781000377965.
2. Khilnani AK. Potential of Large Language Model (ChatGPT) in Constructing Multiple Choice Questions [consultado 13 Sep 2023]. Disponible en: https://zenodo.org/record/7751267. 20 de marzo de 2023.
3. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ. 2024;24(1):354.
4. Masters K, Benjamin J, Agrawal A, MacNeill H, Pillow MT, Mehta N. Twelve tips on creating and using custom GPTs to enhance health professions education. Med Teach. 2024;1-5.
5. Coşkun Ö, Kıyak YS, Budakoğlu İİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach. 2024;0(0):1–7.
6. Rivera-Rosas CN, Calleja-López JRT, Ruibal-Tavares E, Villanueva-Neri A, Flores-Felix CM, Trujillo-López S. Exploring the potential of ChatGPT to create multiple-choice question exams. Educ Méd [Internet]. 2024;25(4) [consultado 18 May 2024]. Disponible en: http://www.elsevier.es/es-revista-educacion-medica-71-articulo-exploring-potential-chatgpt-create-multiple-choice-S1575181324000457.