

Más allá del puntaje: inteligencia artificial generativa y evaluaciones de alto impacto

Beyond the score: generative artificial intelligence and high-stakes assessments

“¿Qué competencias debemos evaluar y cómo evaluamos el aprendizaje en un mundo en el que la IA puede realizar tareas que antes considerábamos exclusivamente humanas?”

UNESCO, 2025

La inteligencia artificial generativa no ha acabado con el examen estandarizado, más bien ha terminado con su inocencia tecnológica. Hasta hace pocos años, quienes diseñaban una evaluación de alto impacto podían suponer que las preguntas habían sido elaboradas por personas, que las respuestas provenían del sustentante y que la vigilancia excluía los apoyos no autorizados. Hoy, la misma tecnología puede generar y resolver reactivos, explicar el razonamiento, producir variantes, calificar respuestas y señalar conductas presuntamente anómalas. El tema ya no es solamente cómo impedir que alguien “use ChatGPT”, sino qué inferencias sobre su competencia siguen siendo válidas, justas y defendibles.

En las profesiones de la salud, la respuesta importa de manera especial. Los exámenes de admisión, progresión, egreso, licenciamiento y certificación distribuyen oportunidades, habilitan trayectorias

profesionales y contribuyen a proteger a los pacientes. Su legitimidad descansa en una promesa pública: que el puntaje representa conocimientos y capacidades relevantes del sustentante, no su acceso a una suscripción, un segundo dispositivo, una red de ayuda o un sistema de vigilancia imperfecto. Al debilitarse esa promesa, el problema deja de ser solo académico; también es ético, institucional y social.

Los primeros estudios que mostraron que un modelo de lenguaje podía alcanzar resultados cercanos al nivel aprobatorio en preguntas liberadas del *United States Medical Licensing Examination* causaron comprensible inquietud.¹ Sin embargo, contestar reactivos escritos no equivale a ejercer con competencia una profesión clínica: no demuestra responsabilidad, empatía, destreza procedimental, comunicación con pacientes, trabajo en equipo ni capacidad para actuar bajo incertidumbre real. Más bien demuestra que muchos reactivos en formatos tradicionales pueden ser resueltos por sistemas que reconocen patrones lingüísticos y sintetizan información. La inteligencia artificial (IA), además de desafiar la seguridad del examen, expone sus límites como representación de la competencia profesional.

Este es un artículo Open Access bajo la licencia CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Esto no vuelve irrelevantes a los exámenes estandarizados. Estos instrumentos conservan ventajas que no son sencillas de reemplazar: muestreo amplio, condiciones comparables, puntuación reproducible, eficiencia a gran escala y estándares comunes. Son una pieza importante de los sistemas de evaluación, pero no deben confundirse con el sistema completo. El reto consiste en conservar sus fortalezas y evitar que una puntuación aparentemente precisa oculte incertidumbres sobre quién respondió, con qué apoyos y qué atributo fue realmente medido.

La IA generativa actualmente puede intervenir en todo el ciclo de vida del examen. Antes de la aplicación puede funcionar como tutor y elaborar simuladores, pero también facilitar la reconstrucción y difusión de reactivos expuestos. Durante una aplicación remota puede proporcionar respuestas mediante otro dispositivo fuera del campo de la cámara. Después puede apoyar la calificación, el establecimiento de puntos de corte, el análisis forense y la detección de irregularidades. En la construcción del instrumento permite producir borradores de casos, opciones y versiones con rapidez.² Tratarla solo como recurso para hacer trampa conduce a respuestas parciales, ya que simultáneamente puede ser objeto de evaluación, herramienta de producción, apoyo para el sustentante y riesgo para la integridad.

El desafío más importante es mantener la validez de constructo. Los profesionales de la salud cada vez usan más a la IA para buscar información, interpretar datos y apoyar decisiones, por lo que prohibir estas herramientas produciría evaluaciones cada vez menos auténticas. Pero permitir las siempre también sería un error, ya que nadie puede verificar una recomendación algorítmica peligrosa si carece del conocimiento fundamental para reconocerla. Algunas capacidades deben demostrarse de manera independiente; otras deben evaluarse con IA disponible, precisamente para observar si el futuro profesional formula buenas preguntas, contrasta fuentes, identifica errores, protege datos y asume la decisión final. La política adecuada debe ser una definición explícita del constructo y del tipo de ayuda compatible con él.

Otro desafío importante es la autenticidad de la respuesta. Los exámenes a distancia desde el domicilio del sustentante amplían el acceso y reducen traslados, pero ocurren en un ambiente que la institución no controla plenamente. La supervisión remota con algoritmos y observadores humanos puede disuadir ciertas conductas, aunque no convierte una habitación particular en un centro seguro. Tampoco es neutral, se han documentado diferencias en el reconocimiento facial y en la frecuencia con que algunos sistemas marcan como sospechosas a personas según características como el tono de piel.³ Más vigilancia no significa más validez, ya que se pueden generar falsos positivos, ansiedad, invasión de privacidad y barreras para quienes tienen discapacidad, conectividad inestable o condiciones domésticas desfavorables. Una alerta automatizada debe iniciar una revisión, nunca sustituir el juicio ni constituir por sí misma un veredicto.

Un tercer desafío es la calidad de los reactivos producidos con IA. La promesa es atractiva: ampliar bancos, reducir costos y crear versiones con rapidez. Sin embargo, la evidencia aconseja tener prudencia. En un examen de medicina de alto impacto, los reactivos generados por un modelo requirieron menos tiempo de producción, pero fueron más fáciles, incluyeron más problemas de exactitud o pertinencia y tendieron a explorar procesos cognitivos de menor complejidad que los elaborados por expertos.⁴ La IA puede ser una excelente asistente editorial, pero todavía no debe ser autora autónoma de decisiones que afectan carreras profesionales. Todo reactivo operativo necesita alineación con la tabla de especificaciones, revisión de contenido y sesgo, pilotaje y evidencia psicométrica.

¿Qué podemos hacer? Una medida es clasificar las evaluaciones según el papel autorizado de la IA. En unas deberá demostrarse competencia independiente y la herramienta estará prohibida; en otras podrá permitirse de manera limitada y declarada; y en otras su uso será parte del desempeño esperado. En este último caso no bastará una respuesta correcta: habrá que valorar cómo se interrogó al sistema, qué datos se compartieron, cómo se verificó la información, qué incertidumbres se reconocieron

y por qué se aceptó o rechazó la recomendación. La alfabetización en IA debe convertirse en competencia evaluable, no en permiso indiscriminado para delegar el razonamiento.

Otra acción es valorar la dependencia en una sola prueba. Un examen estandarizado puede valorar un dominio amplio de conocimientos, pero las decisiones de mayor consecuencia idealmente deben integrar fuentes complementarias: exámenes clínicos objetivos estructurados, simulación, observación en el lugar de trabajo, exámenes orales estructurados, portafolios y seguimiento longitudinal.² La combinación reduce el riesgo de que una modalidad, una filtración o una tecnología distorsione por completo la decisión. La estandarización aporta comparabilidad, pero la evaluación auténtica aporta contexto.

Otra medida es aplicar seguridad proporcional al riesgo. Para decisiones irreversibles o de gran impacto, los centros supervisados continúan siendo la opción más defendible cuando son viables. Si la modalidad remota es necesaria por razones de acceso, debe validarse en esa población y acompañarse de autenticación robusta, control de dispositivos, protocolos ante interrupciones, protección de datos y alternativas razonables. La seguridad requiere múltiples capas: bancos suficientemente amplios, formas equivalentes, control de exposición, monitoreo de filtraciones y análisis forense de respuestas y tiempos. Sin embargo, multiplicar versiones no resuelve por sí solo el problema. Cada forma debe respetar la misma tabla de especificaciones y demostrar dificultad y precisión comparables mediante pilotaje, calibración y equiparación.

Otra medida es establecer una colaboración humano-IA con responsabilidades claras. Los modelos pueden proponer casos, distractores y variantes, revisar claridad y apoyar traducciones. Los expertos clínicos y en evaluación deben seleccionar, corregir, validar y responder por el resultado. Es importante registrar el modelo y su versión, las instrucciones utilizadas, las modificaciones humanas y el desempeño posterior de cada reactivo. En exámenes de alto impacto la trazabilidad es parte de la evidencia de calidad.

Una acción más es gobernar la detección de irregularidades con debido proceso. Un puntaje algorítmico de “sospecha” no debería anular automáticamente un examen. Las decisiones deben integrar varias fuentes independientes, aplicar criterios definidos antes de la administración, ser revisadas por personas capacitadas y ofrecer al sustentante información suficiente y una vía de apelación. También deben auditarse los resultados por subgrupos para identificar efectos diferenciales. La institución, no el proveedor tecnológico, debe conservar la responsabilidad por la decisión.

Una medida más es proteger la privacidad y transparentar las reglas. Antes del examen, los participantes deben saber qué uso de IA está permitido, qué datos serán capturados, quién tendrá acceso, cuánto tiempo se conservarán y cómo se investigará una posible irregularidad. Debe recolectarse solo la información necesaria, con seguridad verificable y contratos que impidan usos secundarios. Las directrices internacionales para evaluación basada en tecnología insisten en considerar de manera conjunta validez, equidad, accesibilidad, seguridad y privacidad, así como contar con planes escritos de prevención, disuasión, detección, respuesta y apelación.⁵

Finalmente, se necesita gobernanza institucional permanente. Un grupo que incluya especialistas clínicos, educadores, psicometristas, seguridad informática, protección de datos, asuntos jurídicos y representantes de sustentantes, pacientes y sociedad debe revisar riesgos, autorizar usos, evaluar proveedores y explicitar sus criterios. Los escenarios de riesgo deben probarse antes de cada ciclo, los incidentes analizarse con transparencia y las políticas actualizarse conforme cambien los modelos. No podemos certificar competencia mediante procesos que la propia institución no comprende.

La inteligencia artificial generativa no obliga a elegir entre conservar intacto el examen tradicional o abolirlo; más bien, debe conducirnos a diseñar mejores programas de evaluación, distinguir la competencia humana independiente de la competencia mediada por tecnología y colocar cada herramienta en el lugar que le corresponde. La seguridad sin

validez produce exámenes blindados pero vacíos, la innovación sin gobernanza produce decisiones rápidas pero indefendibles. En las profesiones de la salud, el objetivo no es demostrar que podemos

vigilar más y automatizar más, sino que podemos tomar decisiones más justas, sustentadas y dignas de la confianza de estudiantes, instituciones, pacientes y sociedad. 🔍



Melchor Sánchez Mendiola

EDITOR

Facultad de Medicina, UNAM

REFERENCIAS

1. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198.
2. Masters K, MacNeil H, Benjamin J, Carver T, Nemethy K, Valanci-Aroesty S, et al. Artificial intelligence in health professions education assessment: AMEE Guide No. 178. Med Teach. 2025;47(9):1410-24. doi:10.1080/0142159X.2024.2445037.
3. Yoder-Himes DR, Asif A, Kinney K, Brandt TJ, Cecil RE, Himes PR, et al. Racial, skin tone, and sex disparities in automated proctoring software. Front Educ. 2022;7:881449. doi:10.3389/feduc.2022.881449.
4. Law AKK, So J, Lui CT, Choi YF, Cheung KH, Hung KKC, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ. 2025;25(1):208. doi:10.1186/s12909-025-06796-6.
5. International Test Commission, Association of Test Publishers. Guidelines for technology-based assessment. Version 1.1. Washington (DC): Association of Test Publishers; 2025. <https://www.intestcom.org/upload/media-library/tba-guidelines-ver-11-july-2025-1754044782Z3vV9.pdf>