

Desarrollo y evidencia de validez de contenido de un instrumento para evaluar la percepción estudiantil sobre una paciente virtual potenciada con inteligencia artificial.

Development and Content Validity Evidence of an Instrument to Assess Students' Perceptions of an AI-Enhanced Virtual Patient.

Carlos Alberto Soto-Aguilera¹, Armando Ortiz-Montalvo², Antonio Cerritos³, Ana Carolina Sepúlveda Vildósola⁴, Juan Andrés Trejo-Mejía⁵, Ana Ivette Mondragón Pineda⁶, Serafín Castañeda Cedeño⁷, Miguel Serrano-Reyes⁸, Florina Gatica-Lara⁹*

Facultad de Medicina, Universidad Nacional Autónoma de México, Ciudad de México, México,

¹ dr.carsot.sem.evaluacion@facmed.unam.mx, <https://orcid.org/0000-0003-1104-2962>,

² aortizm@facmed.unam.mx, <https://orcid.org/0000-0003-3548-1513>

³ drcerritos@facmed.unam.mx, <https://orcid.org/0000-0001-9346-8793>

⁴ anacsepulvedav@facmed.unam.mx, <https://orcid.org/0000-0003-4711-3945>

⁵ andretreme@facmed.unam.mx, <https://orcid.org/0000-0002-0680-6836>

⁶ dra.anaivette.mondragon@facmed.unam.mx, <https://orcid.org/0000-0001-8809-5179>

⁹ floreaga@facmed.unam.mx, <https://orcid.org/0009-0008-7671-3281>

⁷ Facultad de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México, México,

sera_castaneda@comunidad.unam.mx, <https://orcid.org/0000-0002-3252-7953>,

⁸ Facultad de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México, México,

mserrano@ifc.unam.mx, <https://orcid.org/0000-0003-1190-5801>

*Correspondencia: Florina Gatica-Lara. E-mail: floreaga@facmed.unam.mx

Recibido: 28/5/26; Aceptado: 26/6/26; Publicado: 28/6/26

Resumen.

La incorporación de pacientes virtuales potenciados con inteligencia artificial (PV-IA) en educación médica requiere instrumentos que permitan evaluar de manera rigurosa la percepción estudiantil sobre su utilidad, facilidad de uso, realismo y valor formativo. Se desarrolló un cuestionario orientado a valorar la experiencia del estudiantado con el uso de una paciente virtual denominada María V., diseñada con materiales de un ECOE. **Objetivo:** Analizar la evidencia inicial de validez basada en el contenido de un instrumento diseñado para evaluar la percepción estudiantil sobre una PV-IA. **Métodos:** Se realizó un estudio metodológico de validez de contenido. El instrumento integró 23 ítems distribuidos en dimensiones teóricas vinculadas con el Modelo de Aceptación Tecnológica y la educación médica basada en simulación: valor educativo, calidad de la interacción clínica, usabilidad y funcionamiento técnico, realismo y caracterización, satisfacción general y preguntas abiertas. Trece jueces expertos evaluaron la relevancia de cada ítem mediante una escala tipo Likert de cuatro puntos. Se calcularon el índice de validez de contenido por ítem, el índice de validez de contenido a nivel de escala mediante el promedio de los ítems, el índice de acuerdo universal y el kappa modificado. **Resultados:** Los valores del índice de validez de contenido por ítem oscilaron entre 0.69 y 1.00. La mayoría de los ítems alcanzó valores iguales o superiores al punto de corte recomendado. El índice de validez de contenido promedio de la escala fue de 0.916 y el índice de acuerdo universal fue de 0.348. Los valores de kappa modificado oscilaron entre 0.66 y 1.00. Con base en los resultados cuantitativos y las observaciones cualitativas de los jueces, se eliminaron los ítems 6, 9 y 14, y se realizaron ajustes de redacción en los ítems 1, 10, 16 y 18. El cuestionario de percepción sobre la paciente virtual María V. aporta evidencia inicial favorable de validez basada en el contenido. La depuración del instrumento permitió obtener una versión final

de 20 ítems, sustentada no solo en criterios estadísticos, sino también en la coherencia conceptual y funcional con las características actuales de la paciente virtual.

Palabras clave: paciente virtual; inteligencia artificial; educación médica; validez de contenido; simulación clínica; percepción estudiantil.

Abstract.

The incorporation of artificial intelligence-enhanced virtual patients (AI-VP) in medical education requires instruments that rigorously assess students' perceptions of their usefulness, ease of use, realism and educational value. A questionnaire was developed to assess students' experience with a virtual patient named María V., designed based on OSCE materials. To analyze the initial evidence of content validity of an instrument designed to assess students' perceptions of an artificial intelligence-enhanced virtual patient. **Methods:** A methodological content validity study was conducted. The initial instrument consisted of 23 items distributed across theoretically defined dimensions linked to the Technology Acceptance Model and simulation-based medical education: educational value, quality of clinical interaction, usability and technical functioning, realism and characterization, overall satisfaction and open-ended questions. Thirteen expert judges evaluated the relevance of each item using a four-point Likert-type scale. The item-level content validity index, scale-level content validity index based on the average method, universal agreement index and modified kappa were calculated. **Results:** Item-level content validity index values ranged from 0.69 to 1.00. Most items reached values equal to or above the recommended cut-off point. The scale-level content validity index based on the average method was 0.916, and the universal agreement index was 0.348. Modified kappa values ranged from 0.66 to 1.00. Based on quantitative results and the judges' qualitative comments, items 6, 9 and 14 were removed, and adjustments were made to items 1, 10, 16 and 18. The questionnaire on students' perceptions of the virtual patient María V. provides favorable initial evidence of content validity. The refinement of the instrument resulted in a final 20-item version, supported not only by statistical criteria but also by conceptual and functional coherence with the current characteristics of the virtual patient.

Keywords: virtual patient; artificial intelligence; medical education; content validity; clinical simulation; student perception.

1. Introducción

Desde la década de 1960 se ha documentado la evolución conceptual y metodológica de los pacientes estandarizados en el ámbito de la educación médica. En sus inicios, estos actores fueron denominados pacientes programados; posteriormente, el término evolucionó hacia pacientes simulados y, más adelante, hacia pacientes estandarizados, lo que refleja un refinamiento progresivo en su uso educativo y para la evaluación (1-2). En la actualidad, con la incorporación de la inteligencia artificial y el desarrollo de modelos de lenguaje avanzados, emerge la posibilidad de un nuevo cambio de paradigma: la transición de pacientes estandarizados humanos hacia pacientes potenciados por inteligencia artificial (PV-IA), capaces de emitir información clínica estructurada de manera consistente durante los procesos de evaluación o formación (3). Este avance tecnológico permite replantear el papel de los pacientes estandarizados dentro del marco de la evaluación en educación médica, particularmente en evaluaciones de altas consecuencias, al posibilitar la simulación de problemas clínicos de manera controlada, reproducible y escalable. En este sentido, la conceptualización de los pacientes estandarizados ha sido ampliamente discutida en la literatura. Geoff Norman introdujo el término "paciente estandarizado" con el propósito de sustituir la denominación de paciente simulado, enfatizando la necesidad de mantener constante la información clínica proporcionada al estudiantado, a fin de garantizar la estandarización del encuentro clínico. Por su parte, Barrows amplió esta definición al señalar que el término abarca

tanto a individuos sanos capacitados para representar un caso clínico como a pacientes reales que presentan su propia enfermedad de manera estandarizada con fines educativos y de evaluación (1, 4). Una de las principales ventajas de los pacientes estandarizados, tanto humanos como virtuales, radica en su disponibilidad para ser utilizados en diversos escenarios, contextos y momentos, lo que favorece la estandarización, la equidad y la comparabilidad en la evaluación del desempeño clínico (1). En este contexto, la incorporación de PV-IA representa una extensión lógica y prometedora de este enfoque, con implicaciones relevantes para el diseño, la implementación y la innovación de los procesos de evaluación en educación médica.

No obstante, la introducción de nuevas tecnologías en los procesos de evaluación exige garantizar que las inferencias derivadas de los instrumentos utilizados sean válidas y sustentadas en evidencia. En la educación médica contemporánea, la validez no se concibe como una propiedad inherente del instrumento, sino como el grado en que la evidencia y la teoría respaldan las interpretaciones y usos de las puntuaciones obtenidas (5). En esta línea, el enfoque argumentativo de la validez propuesto por Kane (6-7) plantea que la evaluación debe sustentarse en una cadena de inferencias, desde la observación del desempeño hasta la toma de decisiones, que requieren ser respaldadas mediante distintas fuentes de evidencia. En este sentido, diversos autores han enfatizado la necesidad de integrar múltiples fuentes de evidencia de validez en el desarrollo de instrumentos en educación médica, particularmente en contextos de evaluación del desempeño clínico (8). Asimismo, en escenarios de evaluación de altas consecuencias, resulta fundamental asegurar la calidad de la medición para sustentar decisiones educativas y profesionales (9). Dentro de este marco, la validez de contenido constituye una fuente fundamental de evidencia, particularmente en etapas iniciales del desarrollo de instrumentos, ya que permite establecer la relevancia, representatividad y claridad de los ítems en relación con el constructo que se pretende medir. Sin embargo, diversos estudios han señalado limitaciones en el uso aislado de ciertos índices tradicionales de validez de contenido, como el CVI, especialmente en lo referente al acuerdo por azar y la estimación de la precisión de las valoraciones. En este sentido, se ha propuesto complementar estos análisis mediante el estadístico kappa modificado, el cual permite una evaluación más robusta de la validez de contenido. De manera complementaria, la evaluación del desempeño clínico se concibe actualmente como un proceso integrado que requiere múltiples fuentes de información, en lo que se ha denominado evaluación programática, donde distintos instrumentos contribuyen de manera conjunta a la toma de decisiones (10).

A pesar del creciente desarrollo de PV-IA, la evidencia sobre la validez de los instrumentos utilizados para evaluar la percepción estudiantil sobre estas herramientas sigue siendo limitada. En particular, existe una necesidad de contar con instrumentos que aporten evidencia inicial de validez y permitan interpretar adecuadamente las percepciones del estudiantado, así como sustentar su incorporación progresiva en procesos de formación y evaluación clínica. El presente estudio se enmarca en un trabajo de colaboración académica entre la Facultad de Medicina y la Facultad de Ingeniería de la Universidad Nacional Autónoma de México, en el cual se desarrolló un prototipo de PV-IA, denominada María V. Este sistema se basa en la adaptación de un libreto sintético previamente utilizado en un Examen Clínico Objetivo Estructurado (ECOE), lo que permite la interacción verbal entre estudiantes y la paciente virtual, quien responde conforme a una estructura clínica predefinida.

El modelo de aceptación tecnológica (Technology Acceptance Model, TAM) fue adoptado como referente teórico para la construcción de un instrumento orientado a medir la percepción estudiantil sobre esta innovación. Este modelo permite comprender los factores que influyen en la adopción de nuevas tecnologías, particularmente a través de las dimensiones de utilidad percibida y facilidad de uso (11-12). No obstante, en coherencia con los marcos contemporáneos de validez, la construcción y evaluación del instrumento requiere generar evidencia que respalde la interpretación de sus resultados en el contexto educativo. En este contexto, la aportación original

del presente estudio radica en el desarrollo de un instrumento específico para evaluar la percepción estudiantil sobre una PV-IA en un escenario de simulación clínica. A diferencia de aproximaciones centradas únicamente en la aceptación tecnológica, el instrumento integra dimensiones vinculadas con el valor educativo, la calidad de la interacción clínica, la usabilidad, el funcionamiento técnico, el realismo y la satisfacción global. Además, el estudio documenta un proceso sistemático de validez basada en el contenido mediante jueces expertos, I-CVI, S-CVI/Ave, S-CVI/UA y kappa modificado, lo que permite aportar una ruta metodológica inicial para futuras investigaciones sobre tecnologías emergentes en educación médica.

Por ello, el objetivo del presente estudio fue analizar la validez de contenido de un instrumento diseñado para evaluar la percepción estudiantil sobre una PV-IA, mediante el uso combinado del índice de validez de contenido (CVI) y el kappa modificado.

2. Métodos

2.1 Diseño del estudio

Se llevó a cabo un estudio metodológico de validez de contenido, orientado al desarrollo y evaluación inicial de un instrumento de percepción estudiantil, siguiendo la metodología propuesta por Lynn (13). Este enfoque permite estimar el grado en que los ítems de un instrumento representan adecuadamente el constructo de interés mediante la valoración de jueces expertos. El estudio se fundamenta en el marco argumentativo de validez propuesto por Kane (7), en el cual la validez se concibe como la solidez de las inferencias derivadas de las puntuaciones obtenidas. En este contexto, el presente estudio aporta evidencia inicial para sustentar la inferencia de puntuación, específicamente en lo relativo a la relevancia, representatividad y congruencia conceptual de los ítems que conforman el instrumento. Dado que esta fase se centró en la valoración por jueces expertos, las inferencias de generalización, extrapolación e implicación deberán explorarse en estudios posteriores mediante la aplicación del instrumento a población estudiantil, así como mediante análisis de estructura interna, consistencia interna y relaciones con otras variables. Con base en este enfoque, la validez basada en el contenido se analiza como una fuente inicial de evidencia para sustentar la inferencia de puntuación, al evaluar si los ítems representan de manera pertinente y suficiente el constructo que se pretende medir.

2.2 Definición operacional del constructo

Para los fines de este estudio, la percepción estudiantil sobre la PV-IA se definió operacionalmente como: “la valoración que realiza el alumnado acerca de la utilidad educativa, facilidad de uso, calidad de la interacción clínica, realismo, funcionamiento técnico y satisfacción global derivados de su experiencia con una paciente virtual en un escenario de simulación clínica”. Esta definición integra los componentes centrales del TAM, particularmente la utilidad percibida y la facilidad de uso percibida, considerados determinantes en la aceptación de una tecnología (14). Asimismo, incorpora dimensiones propias de la educación médica basada en simulación y del uso de pacientes virtuales, donde la interacción clínica, el realismo del escenario, la fidelidad percibida, la usabilidad y el valor educativo constituyen elementos relevantes para valorar la experiencia formativa del estudiantado (15-16).

Las dimensiones incluidas en el instrumento se concibieron como dominios teóricos de contenido y no como factores latentes empíricamente confirmados. Esta decisión responde al alcance del presente estudio, cuyo propósito fue aportar evidencia inicial de validez basada en el contenido mediante juicio de expertos. En este sentido, la estructura propuesta permite organizar conceptualmente los ítems con base en el TAM y en elementos propios de la educación médica basada en simulación; sin embargo, no constituye aún un modelo de medición validado mediante análisis de estructura interna. Por tanto, la posible organización factorial del instrumento deberá ser

evaluada en estudios posteriores con población estudiantil. En este sentido, la percepción estudiantil no se limita a la aceptación tecnológica de la herramienta, sino que comprende una valoración educativa integral sobre su potencial para favorecer el razonamiento clínico, la comunicación médico-paciente y la preparación para escenarios de evaluación clínica. Con el propósito de explicitar la correspondencia entre el constructo, las dimensiones teóricas y los ítems del cuestionario, se elaboró una matriz de especificaciones del instrumento (tabla 1). Esta matriz permitió organizar los ítems de acuerdo con los dominios conceptuales derivados del TAM y de la educación médica basada en simulación, así como documentar la relación entre cada dimensión, sus indicadores y los ítems correspondientes. Asimismo, constituyó un mecanismo para verificar la representatividad del constructo y orientar el proceso de validación de contenido, al proporcionar evidencia de que los reactivos incluidos cubrían de manera adecuada los aspectos que se pretendían evaluar.

Tabla 1. Matriz de especificaciones del cuestionario de percepción sobre la paciente virtual María V.

Dimensión teórica	Definición conceptual	Fundamento teórico	Ítems iniciales	Ítems finales
Valor educativo	Valoración del estudiantado sobre la contribución de la PV-IA al aprendizaje clínico, el razonamiento clínico y la preparación para escenarios de evaluación clínica.	TAM, particularmente utilidad percibida; educación médica basada en simulación y pacientes virtuales.	1, 2	1, 2
Calidad de la interacción clínica	Valoración de la pertinencia, claridad, suficiencia y utilidad de las respuestas generadas por la paciente virtual durante el interrogatorio clínico.	Simulación clínica, entrevista médica, comunicación médico-paciente y fidelidad funcional de pacientes virtuales.	3, 4, 5, 6, 7	3, 4, 5, 7
Usabilidad y funcionamiento técnico	Valoración de la facilidad de uso, estabilidad, tiempo de respuesta y manejo operativo del sistema durante la interacción con la PV-IA.	TAM, especialmente facilidad de uso percibida; usabilidad de tecnologías educativas.	8, 9, 10, 11, 12	8, 10, 11, 12
Realismo y caracterización	Valoración del grado en que los elementos visuales, auditivos y de caracterización de la paciente virtual favorecen una experiencia de simulación clínica verosímil.	Educación médica basada en simulación, fidelidad percibida, realismo del escenario y diseño de pacientes virtuales.	13, 14, 15, 16, 17, 18	13, 15, 16, 17, 18
Satisfacción general	Valoración global de la experiencia de uso de la PV-IA y disposición del estudiantado para recomendarla en futuros entrenamientos clínicos.	Aceptación tecnológica, experiencia del usuario y percepción global de utilidad educativa.	19, 20	19, 20

Preguntas abiertas complementarias	Exploración cualitativa de aspectos útiles, áreas de mejora y percepción sobre el papel de la PV-IA frente a pacientes estandarizados humanos.	Evaluación cualitativa de experiencias educativas y mejora continua de herramientas de simulación.	21, 22, 23	21, 22, 23
--	---	--	------------	------------

2.3 Diseño de PV-IA María V

La paciente virtual María V fue diseñada como una aplicación web que permite al estudiantado interactuar con ella directamente desde un navegador (17). Su arquitectura se desarrolló mediante el framework Angular y utiliza servicios en la nube de *Amazon Web Services* (AWS) para garantizar estabilidad, rapidez de respuesta y escalabilidad. La interacción se realiza mediante voz, tanto para formular preguntas como para recibir las respuestas de la paciente, con el propósito de aproximarse a la dinámica de una entrevista clínica. El audio emitido por el estudiantado es reconocido automáticamente y transformado en texto mediante servicios de reconocimiento de voz. Posteriormente, dicho texto es procesado por una función AWS Lambda, que se comunica con un modelo de lenguaje de gran escala para generar la respuesta de María V. Esta respuesta se convierte nuevamente en audio mediante servicios de síntesis de voz, permitiendo una interacción verbal continua. A diferencia de otras aplicaciones conversacionales, la comunicación con María V se organiza mediante botones que delimitan los turnos de conversación, lo que permite al estudiantado formular preguntas de manera deliberada y estructurada, como ocurre durante una entrevista clínica.

La información del caso clínico, incluyendo antecedentes, motivo de consulta, exploración física y estudios complementarios, así como las instrucciones que regulan el comportamiento de María V, se almacenan en una base de datos en línea. Esta estructura facilita la incorporación de nuevos casos y la modificación de los existentes sin necesidad de reprogramar la aplicación. Con el propósito de fortalecer la consistencia y pertinencia de las respuestas, se incorporó una capa de agentes encargada de verificar la información generada por el modelo de lenguaje. Estos agentes evalúan que la respuesta sea congruente con los datos del caso, con el rol de la paciente y con la pregunta formulada por el estudiante. Asimismo, filtran información que no sea relevante, que la paciente no debería conocer o que no haya sido solicitada durante la entrevista.

La metainformación del caso, como los hallazgos de exploración física o los resultados de estudios complementarios, se proporciona únicamente cuando el estudiantado realiza la pregunta adecuada o solicita la maniobra correspondiente. Esto permite simular la lógica de una consulta clínica real, en la cual la información se obtiene progresivamente a partir de la interacción entre médico y paciente. El libreto sintético que orienta el comportamiento de María V fue desarrollado con apoyo de expertos e incluye antecedentes personales y familiares, historia de la enfermedad actual, hallazgos de exploración física, resultados de laboratorio y características del comportamiento comunicativo de la paciente durante la entrevista. A partir de este guión se construyeron las instrucciones para el modelo de lenguaje, indicando que María V debía mantenerse en su rol de paciente, responder únicamente con base en la información que una paciente conocería sobre su propio estado, evitar terminología técnica y abstenerse de emitir diagnósticos.

Las características funcionales de María V. orientaron de manera directa la construcción del instrumento. Dado que la herramienta permite una interacción verbal estructurada, la entrega progresiva de información clínica, el uso de botones e íconos para conducir la entrevista, la representación visual mediante un avatar y la generación de respuestas conforme a un caso clínico

predefinido, se consideró necesario incorporar dimensiones relacionadas con la calidad de la interacción clínica, la usabilidad, el funcionamiento técnico, el realismo, la comunicación y el valor educativo de la experiencia. En este sentido, el cuestionario buscó asegurar la correspondencia entre las capacidades reales de la paciente virtual potenciada con inteligencia artificial y los aspectos que el estudiantado podía valorar después de interactuar con ella.

2.4 Desarrollo del instrumento (primera etapa)

En la primera etapa, denominada etapa de desarrollo, participó un grupo académico de la Facultad de Medicina integrado por dos médicos con experiencia en evaluación educativa y en el diseño e implementación del ECOE, así como por dos pasantes de servicio social que interactuaron de manera continua con la paciente virtual María V. durante un periodo de siete meses. La generación inicial de los ítems se llevó a cabo mediante un procedimiento sistemático organizado en cuatro momentos. En primer lugar, se revisaron los referentes conceptuales del TAM y de la educación médica basada en simulación, con el propósito de delimitar las dimensiones teóricas del instrumento. En segundo lugar, se analizaron las características funcionales de María V., incluyendo su modalidad de interacción, la entrega progresiva de información clínica, el funcionamiento técnico, el avatar, la voz y la estructura del caso clínico. En tercer lugar, se revisó el libreto sintético derivado de un ECOE, a fin de identificar los componentes de la experiencia susceptibles de ser valorados por el estudiantado. Finalmente, el grupo académico discutió y depuró los ítems iniciales con base en su pertinencia conceptual, claridad de redacción y correspondencia con las capacidades actuales de la paciente virtual potenciada con inteligencia artificial. A partir de este trabajo colaborativo se diseñó un instrumento compuesto por 23 ítems, distribuidos en distintas dimensiones definidas teóricamente y alineadas con el TAM. Las dimensiones incluidas fueron: valor educativo, calidad de la interacción clínica, usabilidad y funcionamiento técnico, realismo y caracterización, satisfacción general y preguntas abiertas. El instrumento incorporó dos tipos de ítems: de opción múltiple con escala tipo Likert de cuatro puntos (1 = totalmente en desacuerdo; 2 = en desacuerdo; 3 = de acuerdo; 4 = totalmente de acuerdo) y preguntas abiertas orientadas a recabar la percepción del estudiantado sobre el aprendizaje obtenido mediante el uso de la herramienta, posibles áreas de mejora y su potencial para complementar o sustituir el uso de pacientes estandarizados humanos (tabla 2).

Tabla 2. Estructura y dimensiones del “Cuestionario de percepción sobre la paciente virtual María V.”

Sección / Categoría	Ítem	Enunciado
1. Valor educativo	1	El uso de la paciente virtual María V. contribuyó a reforzar mis competencias clínicas para el ECOE.
	2	La experiencia con María V. favoreció mi razonamiento clínico en la resolución de casos.
	3	Las respuestas de la paciente reflejan un interrogatorio clínico real.
2. Calidad de la interacción clínica	4	El nivel de detalle en las respuestas fue suficiente para apoyar el razonamiento clínico.
	5	La comunicación verbal de la paciente fue clara durante el interrogatorio.
	6	El nivel de respuesta emocional de la paciente favoreció la simulación de una entrevista auténtica.
	7	La interacción me permitió desarrollar habilidades de comunicación propias de la relación médico-paciente.
3. Usabilidad y funcionamiento técnico	8	El tiempo de espera entre mis preguntas y las respuestas fue adecuado.
	9	La velocidad de respuesta favoreció la fluidez de la simulación

		clínica.
	10	El sistema me permitió usar de forma sencilla las herramientas disponibles.
	11	El sistema fue estable y confiable durante la interacción (sin fallas técnicas).
	12	La estructura del sistema facilitó concentrarme en el aprendizaje clínico y no en la parte técnica.
4. Realismo y caracterización	13	El entorno visual (fondo) es adecuado para simular un escenario clínico real.
	14	El nivel de detalle del entorno favorece la concentración en la interacción.
	15	El aspecto físico de la paciente virtual resulta realista para un ejercicio de simulación clínica.
	16	Los gestos y expresiones del avatar apoyan la interpretación clínica.
	17	La voz de la paciente virtual fue natural para un escenario clínico.
	18	El tono de voz facilitó la comprensión de la información brindada.
5. Satisfacción general	19	Estoy satisfecho con la experiencia de utilizar la paciente virtual María V.
	20	Recomendaría el uso de este paciente virtual en futuros entrenamientos clínicos.
Preguntas abiertas	21	Desde tu experiencia, ¿qué aspectos del paciente virtual María V. consideras que fueron más útiles para tu aprendizaje clínico?
	22	¿Qué cambios o mejoras sugieres para que el paciente virtual María V. se asemeje más a la interacción con un paciente real en el ECOE?
	23	¿En qué medida consideras que esta herramienta podría complementar o sustituir la práctica con pacientes estandarizados?

2.5 Validación por jueces expertos (segunda etapa)

La segunda etapa, denominada etapa de cuantificación por jueces, consistió en la evaluación del instrumento por un panel de expertos. Para ello, se realizó una invitación a integrantes de la Secretaría de Educación Médica de la Facultad de Medicina y de la Facultad de Ingeniería de la UNAM. La invitación se realizó una semana antes del inicio del proceso de validación y el periodo de evaluación tuvo una duración de dos semanas. El contacto se efectuó mediante mensajería instantánea, en donde se explicó el objetivo del instrumento, su proceso de diseño, el propósito de la validación y el contexto general del proyecto, así como los objetivos de la herramienta educativa. Este proceso tuvo como finalidad obtener evidencia inicial de validez de contenido del instrumento, previa a su aplicación en dos estudios piloto programados para los meses de octubre y noviembre de 2025.

2.6 Características del panel de expertos

Se invitó a un total de 13 expertos, con los siguientes perfiles profesionales: cinco médicos, cuatro psicólogas, tres ingenieros y una pedagoga. El número de jueces se definió de acuerdo con las recomendaciones metodológicas para estudios de validez de contenido, que sugieren un mínimo de 6 a 10 expertos para obtener estimaciones estables del índice de validez de contenido (18). En la tabla 3 se resumen las características del panel de jueces expertos.

Tabla 3. Características del panel de jueces expertos (n = 13).

Variable	Categoría	n (%)
Años de experiencia en	0–5 años	2 (15.4)

educación en ciencias de la salud	6–10 años	4 (30.7)
	11–15 años	1 (7.7)
	16–20 años	2 (15.4)
	21–25 años	2 (15.4)
	>25 años	2 (15.4)
Nombramiento o función en la UNAM	Académico de tiempo completo	3 (23.0)
	Académico de asignatura	1 (7.7)
	Funcionario académico	1 (7.7)
	Académico a tiempo parcial	2 (15.4)
	Académico de tiempo completo y funcionario académico	2 (15.4)
	Académico de asignatura y tiempo parcial	2 (15.4)
	Académico de tiempo completo y de asignatura	1 (7.7)
	Académico de asignatura y funcionario académico	1 (7.7)

2.7 Procedimiento de evaluación

Cada uno de los expertos, de manera cegada e independiente, respondió de forma autoadministrada un formulario electrónico a través de la plataforma Google Forms, en el cual evaluaron el “Cuestionario de percepción sobre la paciente virtual María V”. Los jueces valoraron cada uno de los 23 ítems con el propósito de determinar su pertinencia y adecuación al constructo que se pretendía medir, de acuerdo con los criterios establecidos para la validez de contenido. Para ello, cada ítem fue calificado en términos de relevancia mediante una escala tipo Likert de cuatro puntos: 1 = no es relevante; 2 = poco relevante; 3 = relevante; 4 = muy relevante. Adicionalmente, el formulario incluyó un espacio abierto para que los expertos emitieran recomendaciones y comentarios cualitativos orientados a la mejora del instrumento. Las preguntas abiertas se incorporaron como componentes cualitativos complementarios del instrumento y no como ítems destinados a generar una puntuación cuantitativa. Su valoración por jueces tuvo como propósito estimar la relevancia de su contenido respecto al constructo de percepción estudiantil sobre la paciente virtual potenciada con inteligencia artificial, así como su pertinencia para explorar aspectos relacionados con la utilidad educativa, las oportunidades de mejora de la herramienta y su comparación con pacientes estandarizados humanos. Por ello, estos ítems no serán considerados para el cálculo de puntuaciones de escala, consistencia interna, estructura factorial o modelos de medición en fases posteriores.

2.8 Análisis estadístico

Para el análisis de las respuestas se siguió la metodología propuesta por Polit y colaboradores (18) y se utilizó el software Stata versión 14. Se calcularon los índices de validez de contenido (Content Validity Index, CVI) tanto a nivel de ítem (I-CVI) como a nivel de la escala global (S-CVI), empleando el método de la media aritmética de los I-CVI (S-CVI/Ave). Para el cálculo del I-CVI, se consideró la proporción de jueces que calificaron cada ítem como relevante o muy relevante (valores 3 y 4 de la escala Likert). De acuerdo con las recomendaciones de Polit y cols. (18), los ítems con valores de I-CVI ≥ 0.78 fueron considerados con adecuada validez de contenido. Adicionalmente, se estimó la probabilidad de acuerdo entre jueces, dado que una limitación del CVI es su incapacidad para ajustar el acuerdo atribuible al azar. Para corregir este sesgo, se calculó el estadístico kappa modificado (k^*), propuesto por Wynd y cols. (19) y retomado por Polit y cols. (18), el cual ajusta el nivel de acuerdo considerando la probabilidad de coincidencia por azar en la categoría de relevancia. Los valores de k^* fueron interpretados conforme a los criterios de Fleiss (20)

y Cicchetti y cols. (21) clasificándolos como aceptables, buenos o excelentes según su magnitud. En concordancia con el marco de validez de Kane (7), estos análisis aportan evidencia inicial para sustentar la inferencia de puntuación, al valorar la representatividad, relevancia y congruencia conceptual de los ítems que conforman el instrumento.

2.9 Consideraciones éticas

El presente estudio no fue sometido a evaluación por un comité de ética, debido a que correspondió a una investigación metodológica de validez de contenido centrada en la valoración de un instrumento por jueces expertos, sin intervención educativa directa sobre estudiantes, sin recopilación de datos clínicos, sensibles o identificables de pacientes, y sin implicar riesgos mayores al mínimo para los participantes. La participación de los jueces fue voluntaria, autoadministrada y orientada exclusivamente a valorar la relevancia de los ítems del cuestionario. Asimismo, los participantes firmaron una carta de confidencialidad para resguardar la originalidad del escenario clínico, proteger la propiedad intelectual del material y evitar sesgos prematuros en futuras aplicaciones del instrumento. Las respuestas fueron analizadas de manera agregada, sin presentar resultados individualizados. La revisión por jueces expertos contribuyó a fortalecer el rigor metodológico del instrumento y a generar evidencia inicial de validez basada en el contenido antes de su aplicación en estudios piloto con población estudiantil.

3. Resultados

En relación con los resultados del estudio de validez de contenido, los 23 ítems del “Cuestionario de percepción sobre la paciente virtual María V” fueron evaluados por 13 jueces expertos. Los valores de I-CVI oscilaron entre 0.69 y 1.00. La mayoría de los ítems presentó valores de I-CVI iguales o superiores a 0.78, criterio considerado indicativo de una adecuada validez de contenido cuando participan seis o más expertos. De manera particular, varios ítems alcanzaron un I-CVI de 1.00, lo que refleja acuerdo unánime respecto a su pertinencia. El ítem 14 presentó un valor de I-CVI de 0.69, inferior al punto de corte recomendado, lo que indica una menor concordancia entre los expertos en relación con su relevancia. Asimismo, el ítem 9 mostró un I-CVI de 0.77, valor cercano al criterio mínimo aceptable, por lo que ambos ítems se consideraron susceptibles de revisión con base en las observaciones cualitativas emitidas. En cuanto al ajuste por probabilidad de acuerdo atribuible al azar, los valores del estadístico kappa modificado oscilaron entre 0.66 y 1.00. De acuerdo con los criterios de interpretación propuestos por Fleiss (20) y Cicchetti, la mayoría de los ítems presentó valores de kappa clasificados como buenos a excelentes (tabla 4).

Tabla 4. Índices de validez de contenido por ítem del “Cuestionario de percepción sobre la paciente virtual María V.”

Ítem	N	A	I-CVI	Pc	Kappa modificado (k*)
Ítem 1	13	11	0.85	0.01	0.85
Ítem 2	13	12	0.92	0.00	0.92
Ítem 3	13	13	1.00	0.00	1.00
Ítem 4	13	13	1.00	0.00	1.00
Ítem 5	13	13	1.00	0.00	1.00
Ítem 6	13	11	0.85	0.01	0.85
Ítem 7	13	13	1.00	0.00	1.00
Ítem 8	13	12	0.92	0.00	0.92
Ítem 9	13	10	0.77	0.04	0.76
Ítem 10	13	11	0.85	0.01	0.85
Ítem 11	13	13	1.00	0.00	1.00
Ítem 12	13	13	1.00	0.00	1.00
Ítem 13	13	11	0.85	0.01	0.85

Ítem 14	13	9	0.69	0.09	0.66
Ítem 15	13	12	0.92	0.00	0.92
Ítem 16	13	12	0.92	0.00	0.92
Ítem 17	13	12	0.92	0.00	0.92
Ítem 18	13	11	0.85	0.01	0.85
Ítem 19	13	12	0.92	0.00	0.92
Ítem 20	13	13	1.00	0.00	1.00
Ítem 21	13	12	0.92	0.00	0.92
Ítem 22	13	13	1.00	0.00	1.00
Ítem 23	13	12	0.92	0.00	0.92

Abreviaturas: N = número de jueces; A = número de acuerdos; I-CVI = índice de validez de contenido por ítem; Pc = probabilidad de acuerdo por azar; k* = kappa modificado.

Los ítems con valores de kappa inferiores a 0.75 coincidieron con aquellos que presentaron menores valores de I-CVI. A nivel de escala, el índice de validez de contenido calculado mediante el método de la media aritmética de los I-CVI (S-CVI/Ave) fue de 0.916, valor que supera el criterio recomendado de 0.90 para considerar una adecuada validez de contenido global. Este resultado sugiere que, en conjunto, el instrumento presenta una adecuada representatividad del constructo evaluado. Por el contrario, el índice de validez de contenido calculado mediante el método de acuerdo universal (S-CVI/UA) fue de 0.348. La mayoría de los ítems se mantuvo sin modificaciones al presentar valores adecuados de validez de contenido.

3.1 Análisis cualitativo de los ítems modificados y eliminados

Como complemento de los indicadores cuantitativos de validez de contenido, se analizaron las observaciones cualitativas emitidas por los jueces expertos con el propósito de tomar decisiones sobre la permanencia, modificación o eliminación de los ítems. Este análisis permitió identificar que algunos ítems requerían ajustes de redacción para mejorar su precisión conceptual, mientras que otros debían eliminarse por presentar menor concordancia entre jueces o por no corresponder plenamente con las características funcionales actuales de la PV-IA.

El ítem 1 fue modificado porque su redacción inicial hacía referencia al reforzamiento de “competencias clínicas para el ECOE”, lo que podía incluir componentes amplios no evaluados directamente por la herramienta, como exploración física, toma de decisiones terapéuticas o ejecución procedimental. Por ello, se sustituyó esta formulación por “habilidades clínicas”, con el propósito de delimitar con mayor precisión el alcance formativo de la PV-IA, particularmente en relación con la anamnesis, el razonamiento clínico y la comunicación durante el interrogatorio.

El ítem 10 fue ajustado para precisar la dimensión de usabilidad del sistema. La redacción original hacía referencia de manera general a que el sistema permitía usar de forma sencilla las herramientas disponibles; sin embargo, se consideró necesario especificar con mayor claridad los elementos de interacción. Por ello, la versión final señala que el sistema demostró ser de fácil manejo, permitiendo una interacción sencilla con los botones e íconos disponibles para la realización de las tareas. Este ajuste permitió hacer más explícito el componente operativo de la experiencia de uso.

Respecto a los ítems eliminados, el ítem 6 fue retirado a pesar de haber obtenido valores adecuados de I-CVI y kappa modificado. Esta decisión respondió a un criterio de coherencia conceptual y funcional, ya que el contenido del ítem atribuía a la paciente virtual un nivel de respuesta emocional que no correspondía plenamente con las capacidades actuales de la herramienta. En particular, se consideró que el reactivo podía inducir al estudiantado a valorar una dimensión emocional o afectiva que aún no forma parte del funcionamiento real de la PV-IA. Por

tanto, su eliminación permitió mantener la correspondencia entre los ítems del instrumento, el constructo definido y las características actuales de María V.

El ítem 9 fue eliminado debido a que obtuvo un I-CVI de 0.77 y un kappa modificado de 0.76, valores cercanos al punto de corte mínimo establecido. Aunque la velocidad de respuesta puede influir en la experiencia de interacción, las observaciones cualitativas sugirieron que el ítem podía interpretarse de distintas maneras, al no diferenciar con claridad entre latencia técnica del sistema, fluidez en la conversación con la paciente virtual y continuidad pedagógica de la simulación clínica. Por esta razón, se decidió eliminarlo de la versión final para evitar ambigüedad interpretativa.

El ítem 14 fue eliminado por presentar los valores más bajos de validez de contenido, con un I-CVI de 0.69 y un kappa modificado de 0.66. Estos resultados indicaron menor acuerdo entre los jueces respecto a su relevancia para el constructo evaluado. Además, las observaciones cualitativas sugirieron que el nivel de detalle del entorno visual, aunque puede influir en la concentración y en la percepción de realismo, no constituía un indicador suficientemente central ni específico para valorar la percepción estudiantil sobre la utilidad educativa, clínica y funcional de la PV-IA. Por ello, se decidió retirarlo del instrumento final.

El ítem 16 fue sometido a un ajuste de redacción con el propósito de precisar el papel de los gestos y expresiones del avatar dentro de la experiencia clínica simulada. La modificación buscó evitar que el reactivo sugiriera una interpretación clínica o diagnóstica directa a partir de las expresiones faciales y orientar su sentido hacia el apoyo que estos elementos visuales pueden brindar al realismo, la caracterización de la paciente virtual y la comprensión de la interacción.

El ítem 18 también fue ajustado en su redacción para precisar que el tono de voz debía valorarse en función de su contribución a la comprensión de la información clínica proporcionada por la paciente virtual, y no como una característica estética o aislada del sistema. La modificación buscó fortalecer la correspondencia entre el componente auditivo de la interacción y el constructo de percepción estudiantil sobre la calidad comunicativa y funcional de la PV-IA.

En total, cuatro ítems fueron modificados (1, 10, 16 y 18) y tres fueron eliminados (6, 9 y 14), mientras que los 16 ítems restantes se conservaron sin cambios sustantivos. Como resultado de este proceso, el instrumento final quedó integrado por 20 ítems: 17 ítems con escala tipo Likert y tres preguntas abiertas. La tabla 5 resume exclusivamente los ítems que fueron modificados o eliminados tras el análisis cuantitativo y cualitativo.

Tabla 5. Ítems modificados o eliminados tras el proceso de validación de contenido.

Ítem original	Decisión	Razón principal	Justificación cuantitativa	Justificación cualitativa/conceptual
1	Modificado	Delimitación de la competencia evaluada	I-CVI = 0.85; k* = 0.85	Se sustituyó “competencias clínicas” por “habilidades clínicas” para evitar una referencia demasiado amplia al desempeño clínico general y delimitar mejor el alcance formativo de la PV-IA en el contexto del ECOE.
6	Eliminado	Incongruencia conceptual con las capacidades actuales de la PV-IA	I-CVI = 0.85; k* = 0.85	Aunque obtuvo índices adecuados, se eliminó porque atribuía a la PV-IA un nivel de respuesta emocional que no corresponde plenamente con sus capacidades actuales.
9	Eliminado	Valor cercano	I-CVI = 0.77;	No diferenciaba con claridad entre

		al punto de corte y ambigüedad interpretativa	$k^* = 0.76$	latencia técnica, fluidez en la conversación y continuidad pedagógica de la simulación.
10	Modificado	Precisión de la facilidad de manejo del sistema	I-CVI = 0.85; $k^* = 0.85$	Se precisó la usabilidad del sistema mediante una redacción más específica sobre la interacción sencilla con los botones e íconos disponibles para la realización de las tareas.
14	Eliminado	Baja puntuación en los índices de validez	I-CVI = 0.69; $k^* = 0.66$	El nivel de detalle del entorno visual fue considerado menos central y específico para el constructo evaluado.
16	Modificado	Precisión conceptual del papel de los gestos y expresiones	I-CVI = 0.92; $k^* = 0.92$	Se ajustó la redacción para referirse a “las expresiones del avatar”, evitando atribuir una función diagnóstica o emocional excesiva a los gestos de la paciente virtual.
18	Modificado	Precisión del componente comunicativo auditivo	I-CVI = 0.85; $k^* = 0.85$	Se precisó que el tono de voz debía valorarse por su contribución a la comprensión de la información clínica y a la calidad comunicativa de la interacción.

En conjunto, estos resultados aportan evidencia inicial de validez basada en el contenido, al respaldar la relevancia y representatividad de los ítems como indicadores del constructo de percepción estudiantil sobre la PV-IA. Esta evidencia contribuye de manera preliminar a sustentar la inferencia de puntuación dentro del argumento de validez.

4. Discusión

Los resultados del presente estudio aportan evidencia favorable de validez basada en el contenido para el “Cuestionario de percepción sobre la paciente virtual María V”, tanto a nivel de ítem como de escala. La mayoría de los ítems alcanzó valores de I-CVI iguales o superiores al punto de corte recomendado, lo que sugiere que los expertos consideraron los ítems como relevantes y representativos del constructo relacionado con la percepción estudiantil sobre la PV-IA (18). De manera particular, varios ítems obtuvieron valores de I-CVI de 1.00, reflejando acuerdo unánime respecto a su pertinencia, criterio considerado indicativo de adecuada validez de contenido cuando participan seis o más expertos (13).

Asimismo, la mayoría de los ítems presentaron valores de kappa modificado clasificados como buenos a excelentes, lo que indica un elevado nivel de acuerdo entre los expertos más allá del azar. Este hallazgo fortalece la consistencia de las valoraciones emitidas y respalda la pertinencia de los ítems que conforman el instrumento. Los ítems que presentaron valores de kappa inferiores a 0.75 coincidieron con aquellos que mostraron menores valores de I-CVI, lo que confirma la consistencia entre ambos indicadores y fortalece la decisión metodológica de revisar o eliminar dichos ítems. Desde el marco argumentativo de validez, esta consistencia contribuye a sustentar la inferencia de puntuación, al demostrar que el proceso de evaluación fue sensible para discriminar entre ítems más y menos pertinentes (5-6).

La depuración del instrumento evidenció la necesidad de integrar criterios cuantitativos, cualitativos y conceptuales durante los procesos de validez de contenido. Los ítems 9 y 14 fueron eliminados principalmente por sus menores valores de I-CVI y kappa modificado, así como por

observaciones cualitativas de los jueces relacionadas con ambigüedad interpretativa o menor centralidad respecto al constructo. En contraste, la eliminación del ítem 6 no respondió a un desempeño psicométrico bajo, sino a un criterio de congruencia funcional, dado que hacía referencia a la respuesta emocional de la paciente virtual, una característica que María V. aún no se reproduce de manera consistente. Mantener este ítem habría implicado solicitar al estudiantado valorar una dimensión insuficientemente representada en la experiencia de interacción, lo que podría afectar la interpretación de las puntuaciones. En conjunto, estas decisiones muestran que la permanencia de un ítem no debe depender exclusivamente de puntos de corte estadísticos, sino también de su coherencia conceptual, su adecuación al propósito educativo del instrumento y su correspondencia con las capacidades reales de la tecnología evaluada.

A nivel de escala, el valor obtenido en el S-CVI/Ave superó el criterio recomendado de 0.90, lo que indica una adecuada representatividad global del constructo evaluado y aporta evidencia suficiente de validez de contenido (18). En contraste, el índice calculado mediante el método de acuerdo universal (S-CVI/UA) fue menor, 0.348. Esta diferencia resulta metodológicamente relevante, ya que el S-CVI/Ave refleja el nivel promedio de acuerdo entre los expertos respecto a la relevancia de los ítems, mientras que el S-CVI/UA exige consenso absoluto en cada uno de ellos, por lo que constituye un criterio considerablemente más estricto (18). En instrumentos con un número elevado de ítems y paneles amplios de jueces, como en el presente estudio, la probabilidad de alcanzar acuerdo universal disminuye sustancialmente, aun cuando los ítems individuales sean pertinentes. Por ello, el bajo valor del S-CVI/UA no debe interpretarse como evidencia de insuficiente validez de contenido, sino como reflejo de la naturaleza restrictiva de este indicador. De manera complementaria, permitió identificar los ítems con menor consenso entre los expertos y orientar su revisión, modificación o eliminación durante el proceso de refinamiento del instrumento.

La principal contribución de este estudio es que propone y documenta un proceso metodológico explícito para construir y depurar un instrumento orientado a evaluar la percepción estudiantil sobre una PV-IA; además, este enfoque de validez es relevante porque la rápida incorporación de herramientas basadas en IA en educación médica exige no solo evaluar su factibilidad técnica, sino también contar con instrumentos que permitan interpretar de manera rigurosa la experiencia formativa del estudiantado. En este sentido, el cuestionario desarrollado no se limita a medir aceptación tecnológica, sino que incorpora dimensiones educativas y clínicas propias de la simulación, como la calidad de la interacción, el realismo percibido y el valor formativo de la experiencia.

En comparación con algunos estudios internacionales relacionados con pacientes virtuales en el ECOE, en los que no siempre se describen procedimientos explícitos de validez basada en el contenido, este estudio describe de manera transparente el proceso de construcción, revisión por expertos y depuración del instrumento. Como se documentó en la revisión de la literatura realizada en esta investigación, diversos instrumentos previamente utilizados no reportan procesos sistemáticos de validación de contenido mediante jueces expertos equivalentes al desarrollado en este estudio. En este sentido, el principal aporte de la presente investigación radica en la transparencia, sistematización y rigurosidad metodológica aplicada durante el proceso de validación (22-24).

El rigor metodológico implementado permitió no solo desarrollar una herramienta funcional para evaluar la percepción estudiantil sobre la PV-IA, sino también generar una hoja de ruta metodológica replicable para futuras investigaciones (25). Este aspecto representa una contribución relevante para la educación médica, al establecer un referente para el diseño y validación de métricas orientadas a evaluar experiencias educativas mediadas por inteligencia artificial.

Desde una perspectiva educativa, la incorporación de PV-IA transforma el paradigma del aprendizaje simulado, permitiendo al estudiantado practicar habilidades de anamnesis, razonamiento clínico y comunicación en entornos seguros y con alta fidelidad dialógica (26). Sin embargo, la efectividad pedagógica de estas herramientas depende en gran medida de la aceptación, percepción de utilidad, facilidad de uso y realismo experimentado por los estudiantes; asimismo, de la alineación del constructo que se pretende medir. En este contexto, disponer de instrumentos con evidencia inicial de validez de contenido permite avanzar hacia una evaluación más sistemática de la interacción entre el alumnado y estas tecnologías, contribuyendo a optimizar los escenarios clínicos, identificar riesgos potenciales, como respuestas inexactas generadas por la inteligencia artificial, y fortalecer los procesos de realimentación y aprendizaje.

Por otro lado, desde una perspectiva curricular, los hallazgos de validez de contenido permiten identificar dimensiones con potencial para orientar la incorporación gradual de PV-IA en programas de formación clínica. Los ítems con mayor consenso entre expertos se relacionaron con el valor educativo, la calidad de la interacción clínica, la claridad de la comunicación verbal, el razonamiento clínico y la satisfacción general, dimensiones que pueden vincularse con actividades de entrenamiento para el ECOE, prácticas de anamnesis, comunicación médico-paciente y preparación para escenarios clínicos simulados. En este sentido, el instrumento no solo permite valorar la aceptación de una herramienta tecnológica, sino también generar información útil para el diseño curricular, la mejora de experiencias de simulación clínica y la toma de decisiones sobre la integración progresiva de pacientes virtuales en la educación médica.

Limitaciones del estudio

La validez de contenido constituye únicamente una de las fuentes de evidencia de validez contempladas en los marcos contemporáneos de evaluación (27). En este sentido, los resultados reportados no permiten establecer aún evidencia relacionada con la estructura interna del instrumento, las relaciones con otras variables o la validez de criterio. Asimismo, aunque el panel de expertos estuvo conformado por profesionales con experiencia diversa en educación y ciencias de la salud, la interpretación de los resultados continúa dependiendo del juicio subjetivo de los evaluadores. Por otra parte, el estudio se centró exclusivamente en la evaluación de la percepción estudiantil sobre una PV-IA dentro de un contexto específico, por lo que la generalización de los resultados a otros entornos educativos o tecnológicos debe realizarse con cautela. Otra limitación del presente estudio es que la valoración cuantitativa realizada por los jueces se centró en la relevancia y representatividad de los ítems del instrumento. Si bien los comentarios cualitativos permitieron identificar oportunidades de mejora relacionadas con la claridad, coherencia y congruencia conceptual de los ítems, estas dimensiones no fueron evaluadas mediante escalas independientes que permitieran cuantificar formalmente el grado de acuerdo entre los expertos. En consecuencia, la evidencia obtenida se concentra principalmente en la validez de contenido asociada a la relevancia de los reactivos. Futuras investigaciones podrían fortalecer este proceso incorporando valoraciones diferenciadas de claridad, coherencia y suficiencia. Asimismo, las dimensiones propuestas deben interpretarse como dominios teóricos de contenido y no como factores latentes empíricamente confirmados. En consecuencia, no se recomienda aún interpretar puntuaciones por subescala ni asumir una estructura factorial del instrumento. Para fortalecer el argumento de validez, estudios posteriores deberán evaluar la estructura interna, la consistencia interna, el comportamiento de las dimensiones y su asociación con otras variables mediante la aplicación del instrumento en población estudiantil (5-6).

5. Conclusiones

- Este trabajo aporta una ruta metodológica inicial para el desarrollo de instrumentos orientados a evaluar experiencias educativas mediadas por inteligencia artificial en simulación clínica. Su contribución principal radica en integrar un marco teórico de aceptación tecnológica con dimensiones propias de la educación médica basada en

simulación, así como en documentar un proceso sistemático de revisión por jueces expertos y depuración de ítems. Futuras investigaciones deberán aplicar el instrumento en población estudiantil para aportar nuevas fuentes de evidencia de validez, incluyendo estructura interna, consistencia interna y relaciones con otras variables.

- La eliminación de los ítems 9, 14 y 6 refleja la importancia de considerar no solo criterios estadísticos, sino también la coherencia conceptual y funcional del instrumento durante los procesos de validación. En conjunto, este proceso permitió depurar el cuestionario y fortalecer su pertinencia para su aplicación en estudios piloto posteriores. Los hallazgos del presente estudio aportan evidencia inicial de validez basada en el contenido para el “Cuestionario de percepción sobre la paciente virtual María V”, contribuyendo al desarrollo de herramientas metodológicamente sólidas para evaluar la percepción estudiantil sobre tecnologías emergentes en educación médica.

Financiación: No ha habido financiación.

Declaración de conflicto de interés: Los autores declaran no tener ningún conflicto de interés.

Contribuciones de los autores: Conceptualización: CASA, AOM, ACSV, JATM, SCC y FGL. Metodología: CASA y FGL. Software: SCC y MSR. Validación: AOM, AC, JATM, AIMP, SCC y FGL. Análisis formal: CASA, AC y AOM. Investigación: CASA, MSR y FGL. Recursos: ACSV, SCC y MSR. Curación de datos: CASA. Redacción - borrador original: CASA, AOM, AC, JATM, AIMP, SCC, MSR y FGL. Redacción - revisión y edición: CASA, AOM, AC, ACSV, JATM, AIMP, SCC, MSR y FGL. Visualización: AOM, AC y AIMP. Supervisión: AOM y ACSV. Administración del proyecto: CASA, AOM, ACSV y SCC. Adquisición de fondos: no aplica.

6. Referencias.

1. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Acad Med*. **1993**, 68, 6, 443-451. <https://doi.org/10.1097/00001888-199306000-00002>
2. Cross J, Kayalackakom T, Robinson RE, Vaughans A, Sebastian R, Hood R, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. *JMIR Med Educ*. **2025**, 11. <https://doi.org/10.2196/63353>
3. Zhang B, Liu X, Wang Y, Zhou L, Xie Q, Wang B. Human or LLM as standardized patients? A comparative study for medical education. *arXiv*. **2025**. <https://doi.org/10.48550/arXiv.2511.14783>.
4. Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education. *Cureus*. **2024**, 16, 10. <https://doi.org/10.7759/cureus.71224>
5. Messick S. Validity. In: Linn RL, editor. *Educational measurement*. 3rd ed. New York: Macmillan; **1989**. p. 13-103. <https://eric.ed.gov/?id=ED372105>
6. Kane MT. Validation. In: Brennan RL, editor. *Educational measurement*. 4th ed. Westport (CT): American Council on Education/Praeger; **2006**. p. 17-64. <https://eric.ed.gov/?id=ED493398>
7. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. **2013**, 50, 1. 1-73. <https://doi.org/10.1111/jedm.12000>
8. Carrillo-Avalos BA, Leenen I, Trejo-Mejía JA, Sánchez-Mendiola M. Bridging validity frameworks in assessment: beyond traditional approaches in health professions education. *Teach Learn Med*. **2025**, 37, 2. 229-238. <https://doi.org/10.1080/10401334.2023.2293871>
9. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. **2003**, 37, 9, 830-837. <https://doi.org/10.1046/j.1365-2923.2003.01594.x>
10. Van der Vleuten CPM, Schuwirth LWT, Driessen EW et al. A model for programmatic assessment fit for purpose. *MedTeach*. **2012**, 34, 3. 205-214. <https://doi.org/10.3109/0142159X.2012.652239>
11. Briz-Ponce L, García-Peñalvo FJ. An empirical assessment of a technology acceptance model for applications in medical education. *J Med Syst*. **2015**, 39, 176. <https://doi.org/10.1007/s10916-015-0352-x>
12. Lee JWY, Ong DW, Soh RCC, Rao JP, Bello F. Technology Acceptance Model in medical education: systematic review. *JMIR Med Educ*. **2025**, 11. <https://doi.org/10.2196/67873>
13. Lynn MR. Determination and quantification of content validity. *Nurs Res*. **1986**, 35, 6. 382-385.
14. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MY Q*. **1989**, 13, 3. 319-340. <https://doi.org/10.2307/249008>
15. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. *Acad Med*. **2010**, 85, 10. 1589-1602. <https://doi.org/10.1097/ACM.0b013e3181edfe13>

16. INACSL Standards Committee. Healthcare Simulation Standards of Best Practice®. *Clin Simul Nurs*. **2021**, 58, 1-93. <https://doi.org/10.1016/j.ecns.2021.08.018>
17. Zeng J, Qi W, Shen S et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. *J Med Internet Res*. **2025**, 27. <https://doi.org/10.2196/79091>.
18. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health*. **2007**, 30, 4. 459-467. <https://doi.org/10.1002/nur.20199>
19. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. *West J Nurs Res*. **2003**, 25, 5. 508-518. <https://doi.org/10.1177/0193945903252998>
20. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. **1971**, 76, 5, 378-382. <https://doi.org/10.1037/h0031619>
21. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: application to assessment of adaptive behavior. *Am J Ment Defic*. **1981**, 86, 2. 127-137. <https://pubmed.ncbi.nlm.nih.gov/7315877/>
22. Grevisse C, et al. ECOSBot: a generative AI-based tool designed to simulate both standardized patients and standardized examiners in nephrology-focused OSCEs. *Clin Kidney J*. **2025**, 18, 10. <https://doi.org/10.1093/ckj/sfaf308>
23. Borg A, Schiött J, Ivegren W, Gentline C et al. AI-generated feedback following social robotic virtual patient interactions and medical student performance: nonrandomized quasi-experimental study. *JMIR Med Educ*. **2026**, 12. <https://doi.org/10.2196/90368>
24. Gamble C, Oatham A, Parikh R. Should virtual OSCE teaching replace or complement face-to-face teaching in the post-COVID-19 educational environment. *Cureus*. **2023**, 15, 11. <https://doi.org/10.7759/cureus.49708>
25. Khojasteh L, Karimian Z, Nasiri E, Kafipour R, Farahmandi AY. Artificial intelligence and academic writing questionnaire (AI-AWQ): development and validation among medical students' experiences using exploratory factor analysis. *BMC Med Educ*. **2025**, 25, 1. 1697. <https://doi.org/10.1186/s12909-025-08288-z>
26. Yi Y, Kim KJ. The feasibility of using generative artificial intelligence for history taking in virtual patients. *BMC Res Notes*. **2025**, 18, 1. 80. <https://doi.org/10.1186/s13104-025-07157-8>
27. American Psychological Association, National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington (DC): American Educational Research Association; **2014**. <https://www.apa.org/science/programs/testing/standards>

Copyright



© 2026 University of Murcia. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 Spain License (CC BY-NC-ND) (<http://creativecommons.org/licenses/by/4.0/>).