Revista Española de Educación Médica

Development and content validity evidence of an instrument to evaluate student perception of a virtual patient powered by artificial intelligence.

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

Carlos Alberto Soto-Aguilera[1], Armando Ortiz-Montalvo[2], Antonio Cerritos[3], Ana Carolina Sepúlveda Vildósola[4], Juan Andrés Trejo-Mejía[5], Ana Ivette Mondragón Pineda[6], Serafín Castañeda Cedeño[7], Miguel Serrano - Reyes[8], Florina Gatica-Lara[9*]

Faculty of Medicine, National Autonomous University of Mexico, Mexico City, Mexico,
[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] floregala@facmed.unam.mx, https://orcid.org/0009-0008-7671-3281
[7] Faculty of Engineering, National Autonomous University of Mexico, Mexico City, Mexico, sera_castaneda@comunidad.unam.mx, https://orcid.org/0000-0002-3252-7953,
[8] Faculty of Engineering, National Autonomous University of Mexico, Mexico City, Mexico, mserrano@ifc.unam.mx, https://orcid.org/0000-0003-1190-5801

*Correspondence: Florina Gatica-Lara. E-mail: floregala@facmed.unam.mx

Received: 28/5/26; Accepted: 26/6/26; Published: 28/6/26

Summary.
The incorporation of AI-enhanced virtual patients (AI-VPs) in medical education requires instruments that allow for a rigorous evaluation of student perceptions regarding their usefulness, ease of use, realism, and educational value. A questionnaire was developed to assess students' experience using a virtual patient named María V., designed with materials from a Clinical Evaluation Center (CEC). **Objective:** To analyze the initial content-based validity evidence of an instrument designed to evaluate student perceptions of an AI-VP. **Methods:** A methodological content validity study was conducted. The instrument comprised 23 items distributed across theoretical dimensions related to the Technology Acceptance Model and simulation-based medical education: educational value, quality of clinical interaction, usability and technical functionality, realism and characterization, overall satisfaction, and open-ended questions. Thirteen expert judges evaluated the relevance of each item using a four-point Likert scale. The content validity index for each item, the content validity index at the scale level (using the average of the items), the universal agreement index, and the modified kappa were calculated. **Results:** Content validity index values for each item ranged from 0.69 to 1.00. Most items reached values equal to or higher than the recommended cutoff point. The average content validity index for the scale was 0.916, and the universal agreement index was 0.348. Modified kappa values ranged from 0.66 to 1.00. Based on the quantitative results and the judges' qualitative observations, items 6, 9, and 14 were eliminated, and wording adjustments were made to items 1, 10, 16, and 18. The questionnaire on perceptions of the virtual patient María V. provides favorable initial evidence of content-based validity. The refinement of the instrument allowed us to obtain a final version of 20 items, based not only on

statistical criteria, but also on 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.

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.

---

1. Introduction

Since the 1960s, the conceptual and methodological evolution of standardized patients in medical education has been documented. Initially, these actors were called programmed patients; later, the term evolved to simulated patients and, subsequently, to standardized patients, reflecting a progressive refinement in their educational and assessment use (1-2). Currently, with the incorporation of artificial intelligence and the development of advanced language models, the possibility of a new paradigm shift is emerging: the transition from human standardized patients to AI-powered patients (AI-PV), capable of consistently providing structured clinical information during assessment or training processes (3). This technological advancement allows for a rethinking of the role of standardized patients within the framework of assessment in medical education, particularly in high-stakes assessments, by enabling the controlled, reproducible, and scalable simulation of clinical problems. In this sense, the conceptualization of standardized patients has been widely discussed in the literature. Geoff Norman introduced the term "standardized patient" to replace the term "simulated patient," emphasizing the need to maintain consistency in the clinical information provided to students to ensure the standardization of the clinical encounter. Barrows expanded this definition, noting that the term encompasses both healthy individuals trained to represent a clinical case and real patients who present their own illness in a standardized

manner for educational and assessment purposes (1, 4). One of the main advantages of standardized patients, both human and virtual, lies in their availability for use in diverse scenarios, contexts, and times, which promotes standardization, equity, and comparability in the evaluation of clinical performance (1). In this context, the incorporation of virtual patient-assistance (VPA) represents a logical and promising extension of this approach, with significant implications for the design, implementation, and innovation of assessment processes in medical education.

However, the introduction of new technologies into assessment processes requires ensuring that the inferences derived from the instruments used are valid and supported by evidence. In contemporary medical education, validity is not conceived as an inherent property of the instrument, but rather as the degree to which evidence and theory support the interpretations and uses of the scores obtained (5). Along these lines, Kane's argumentative approach to validity (6-7) posits that assessment should be based on a chain of inferences, from performance observation to decision-making, which require support from various sources of evidence. In this sense, several authors have emphasized the need to integrate multiple sources of validity evidence in the development of instruments in medical education, particularly in clinical performance assessment contexts (8). Likewise, in high-stakes assessment scenarios, ensuring the quality of the measurement is fundamental to supporting educational and professional decisions (9). Within this framework, content validity constitutes a fundamental source of evidence, particularly in the initial stages of instrument development, as it allows for establishing the relevance, representativeness, and clarity of the items in relation to the construct being measured. However, several studies have pointed out limitations in the isolated use of certain traditional content validity indices, such as the CVI, especially regarding chance agreement and the estimation of rating accuracy. In this sense, it has been proposed to complement these analyses with the modified kappa statistic, which allows for a more robust assessment of content validity. Additionally, clinical performance evaluation is currently conceived as an integrated process that requires multiple sources of information, in what has been termed programmatic evaluation, where different instruments contribute jointly to decision-making (10).

Despite the growing development of PV-IA, evidence on the validity of the instruments used to assess student perceptions of these tools remains limited. In particular, there is a need for instruments that provide initial evidence of validity and allow for the proper interpretation of student perceptions, as well as support their progressive incorporation into training and clinical assessment processes. This study is part of an academic collaboration between the Faculty of Medicine and the Faculty of Engineering of the National Autonomous University of Mexico, in which a prototype of PV-AI, called Maria V, was developed. This system is based on the adaptation of a synthetic script previously used in an Objective Structured Clinical Examination (OSCE), which allows verbal interaction between students and the virtual patient, who responds according to a predefined clinical structure.

The Technology Acceptance Model (TAM) was adopted as the theoretical framework for constructing an instrument designed to measure student perceptions of this innovation. This model allows us to understand the factors that influence the adoption of new technologies, particularly through the dimensions of perceived usefulness and ease of use (11-12). However, in accordance with contemporary validity frameworks, the construction and evaluation of the instrument requires generating evidence to support the interpretation of its results in the educational context. In this context, the original contribution of this study lies in the development of a specific instrument to evaluate student perceptions of a PV-AI in a clinical simulation setting. Unlike approaches focused solely on technological acceptance, this instrument integrates dimensions related to educational value, the quality of clinical interaction, usability, technical performance, realism, and overall satisfaction. Furthermore, the study documents a systematic content-based validity process using

expert judges, I-CVI, S-CVI/Ave, S-CVI/UA and modified kappa, which allows for an initial methodological route for future research on emerging technologies in medical education.

Therefore, the objective of this study was to analyze the content validity of an instrument designed to evaluate student perception on a PV-IA, through the combined use of the content validity index (CVI) and the modified kappa.

2. Methods

2.1 Study design

A methodological study of content validity was conducted, focused on the development and initial evaluation of a student perception instrument, following the methodology proposed by Lynn (13). This approach allows for estimating the degree to which the items of an instrument adequately represent the construct of interest through the evaluation of expert judges. This study is based on Kane's (7) argumentative framework for validity, in which validity is conceived as the soundness of inferences derived from the scores obtained. In this context, the present study provides initial evidence to support score inferences, specifically regarding the relevance, representativeness, and conceptual consistency of the items that comprise the instrument. Since this phase focused on expert judge evaluation, inferences of generalization, extrapolation, and implication should be explored in subsequent studies by applying the instrument to a student population, as well as through analyses of internal structure, internal consistency, and relationships with other variables. Based on this approach, content-based validity is analyzed as an initial source of evidence to support score inferences by evaluating whether the items pertinently and sufficiently represent the construct they are intended to measure.

2.2 Operational definition of the construct

For the purposes of this study, student perception of virtual patient-associated simulation (VPS) was operationally defined as: "the students' assessment of the educational utility, ease of use, quality of clinical interaction, realism, technical performance, and overall satisfaction derived from their experience with a virtual patient in a clinical simulation setting." This definition integrates the core components of the Technology Acceptance Model (TAM), particularly perceived utility and perceived ease of use, considered key determinants in the acceptance of a technology (14). It also incorporates dimensions specific to simulation-based medical education and the use of virtual patients, where clinical interaction, scenario realism, perceived fidelity, usability, and educational value are relevant elements for evaluating the students' learning experience (15-16).

The dimensions included in the instrument were conceived as theoretical content domains and not as empirically confirmed latent factors. This decision reflects the scope of the present study, whose purpose was to provide initial evidence of content-based validity through expert judgment. In this sense, the proposed structure allows for the conceptual organization of the items based on the TAM (Theory of Medical Assessment) and elements specific to simulation-based medical education; however, it does not yet constitute a measurement model validated through internal structure analysis. Therefore, the potential factorial organization of the instrument should be evaluated in subsequent studies with student populations. In this sense, student perception is not limited to the technological acceptance of the tool, but rather encompasses a comprehensive educational assessment of its potential to foster clinical reasoning, doctor-patient communication, and preparation for clinical assessment scenarios. To clarify the correspondence between the construct, the theoretical dimensions, and the questionnaire items, an instrument specifications matrix was developed (Table 1). This matrix allowed for the organization of the items according to the conceptual domains derived from the TAM (Theory of Medical Assessment) and simulation-based medical education, as well as documenting the relationship between each dimension, its

indicators, and the corresponding items. Furthermore, it served as a mechanism to verify the representativeness of the construct and guide the content validation process by providing evidence that the included items adequately covered the aspects intended to be evaluated.

Table 1. Specification matrix of the perception questionnaire on the virtual patient Maria V.

Theoretical dimension	Conceptual definition	Theoretical basis	Initial items	Final items
Educational value	Student assessment of PV-IA's contribution to clinical learning, clinical reasoning, and preparation for clinical assessment scenarios.	TAM, particularly perceived usefulness; simulation-based medical education and virtual patients.	1, 2	1, 2
Quality of clinical interaction	Assessment of the relevance, clarity, sufficiency and usefulness of the responses generated by the virtual patient during the clinical interview.	Clinical simulation, medical interview, doctor-patient communication, and functional fidelity of virtual patients.	3, 4, 5, 6, 7	3, 4, 5, 7
Usability and technical operation	Evaluation of ease of use, stability, response time and operational handling of the system during interaction with the PV-AI.	TAM, especially perceived ease of use; usability of educational technologies.	8, 9, 10, 11, 12	8, 10, 11, 12
Realism and characterizati on	Assessment of the degree to which the visual, auditory and characterization elements of the virtual patient contribute to a realistic clinical simulation experience.	Simulation-based medical education, perceived fidelity, scenario realism, and virtual patient design.	13, 14, 15, 16, 17, 18	13, 15, 16, 17, 18
Overall satisfaction	Overall assessment of the PV-IA user experience and students' willingness to recommend it in future clinical training.	Technological acceptance, user experience, and overall perception of educational utility.	19, 20	19, 20
Additional open-ended questions	Qualitative exploration of useful aspects, areas for improvement and perception on the role of PV-IA compared to standardized human patients.	Qualitative evaluation of educational experiences and continuous improvement of simulation tools.	21, 22, 23	21, 22, 23

2.3 PV-IA Design Maria V

The virtual patient Maria V was designed as a web application that allows students to interact with her directly from a browser (17). Its architecture was developed using the Angular framework and utilizes *Amazon Web Services* (AWS) cloud services to ensure stability, responsiveness, and scalability. Interaction is voice-based, both for asking questions and receiving responses from the

patient, in order to approximate the dynamics of a clinical interview. The audio emitted by the students is automatically recognized and transformed into text using speech recognition services. Subsequently, this text is processed by an AWS Lambda function, which communicates with a large-scale language model to generate Maria V's response. This response is then converted back into audio using speech synthesis services, enabling continuous verbal interaction. Unlike other conversational applications, communication with Maria V is organized using buttons that delimit conversation turns, allowing students to ask questions in a deliberate and structured manner, as occurs during a clinical interview.

The clinical case information, including medical history, reason for consultation, physical examination, and complementary studies, as well as the instructions governing Maria V's behavior, are stored in an online database. This structure facilitates the addition of new cases and the modification of existing ones without requiring the application to be reprogrammed. To strengthen the consistency and relevance of the responses, a layer of agents was incorporated to verify the information generated by the language model. These agents evaluate whether the response is congruent with the case data, the patient's role, and the question posed by the student. They also filter out irrelevant information, information the patient should not know, or information not requested during the interview.

Meta-information about the case, such as physical examination findings or the results of additional studies, is provided only when the student asks the appropriate question or requests the corresponding procedure. This allows for the simulation of a real clinical consultation, in which information is obtained progressively through the interaction between doctor and patient. The concise script guiding Maria V's behavior was developed with expert support and includes personal and family history, history of the present illness, physical examination findings, laboratory results, and characteristics of the patient's communication style during the interview. Based on this script, instructions for the language model were developed, indicating that Maria V should remain in her role as a patient, respond only based on information a patient would know about her own condition, avoid technical terminology, and refrain from making diagnoses.

The functional characteristics of María V. directly guided the construction of the instrument. Given that the tool allows for structured verbal interaction, the progressive delivery of clinical information, the use of buttons and icons to conduct the interview, visual representation through an avatar, and the generation of responses according to a predefined clinical case, it was considered necessary to incorporate dimensions related to the quality of the clinical interaction, usability, technical functionality, realism, communication, and the educational value of the experience. In this sense, the questionnaire sought to ensure the correspondence between the actual capabilities of the AI-enhanced virtual patient and the aspects that students could evaluate after interacting with her.

2.4 Instrument Development (First Stage)

In the first stage, called the development stage, an academic group from the Faculty of Medicine participated, made up of two doctors with experience in educational evaluation and in the design and implementation of the OSCE, as well as two social service interns who interacted continuously with the virtual patient Maria V. for a period of seven months. The initial generation of items was carried out through a systematic procedure organized in four stages. First, the conceptual frameworks of the TAM (Theory of Medical Assessment) and simulation-based medical education were reviewed to define the instrument's theoretical dimensions. Second, the functional characteristics of María V. were analyzed, including her interaction mode, the progressive delivery of clinical information, technical functionality, avatar, voice, and clinical case structure. Third, the synthetic script derived from an OSCE (Objective Structured Clinical Examination) was reviewed to identify the components of the experience that could be evaluated by students. Finally, the

academic group discussed and refined the initial items based on their conceptual relevance, clarity of wording, and alignment with the current capabilities of the AI-enhanced virtual patient. From this collaborative work, an instrument composed of 23 items was designed, distributed across different theoretically defined dimensions aligned with the TAM. The dimensions included were: educational value, quality of clinical interaction, usability and technical functioning, realism and characterization, general satisfaction, and open-ended questions. The instrument incorporated two types of items: multiple-choice items with a four-point Likert scale (1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree) and open-ended questions aimed at gathering students' perceptions of the learning obtained through the use of the tool, possible areas for improvement, and its potential to complement or replace the use of standardized human patients (Table 2).

Table 2. Structure and dimensions of "Perception Questionnaire about the virtual patient Maria V."

Section / Category	Item	Statement
1. Educational value	1	The use of the virtual patient Maria V. helped to strengthen my clinical skills for the OSCE.
	2	The experience with Maria V. improved my clinical reasoning in case resolution.
2. Quality of clinical interaction	3	The patient's answers reflect a real clinical questioning.
	4	The level of detail in the responses was sufficient to support the clinical reasoning.
	5	The patient's verbal communication was clear during the questioning.
	6	The patient's level of emotional response facilitated the simulation of an authentic interview.
	7	The interaction allowed me to develop communication skills specific to the doctor-patient relationship.
3. Usability and technical operation	8	The waiting time between my questions and the answers was adequate.
	9	The speed of response favored the fluidity of the clinical simulation.
	10	The system allowed me to easily use the available tools.
	11	The system was stable and reliable during the interaction (without technical failures).
	12	The structure of the system made it easier for me to focus on clinical learning and not on the technical aspects.
4. Realism and characterization	13	The visual environment (background) is suitable for simulating a real clinical scenario.
	14	The level of detail in the environment promotes concentration on the interaction.
	15	The physical appearance of the virtual patient is realistic for a clinical simulation exercise.
	16	The avatar's gestures and expressions support the clinical interpretation.
	17	The virtual patient's voice sounded natural for a clinical setting.
	18	The tone of voice facilitated the understanding of the information provided.
5. Overall satisfaction	19	I am satisfied with the experience of using the virtual patient Maria V.
	20	I would recommend the use of this virtual patient in future clinical training.
Open questions	21	From your experience, what aspects of the virtual patient María V.

	do you consider to have been most useful for your clinical learning?
22	What changes or improvements do you suggest to make the virtual patient Maria V. more closely resemble the interaction with a real patient in the OSCE?
23	To what extent do you think this tool could complement or replace practice with standardized patients?

2.5 Validation by expert judges (second stage)

The second stage, called the quantification stage by judges, consisted of the evaluation of the instrument by a panel of experts. For this purpose, members of the Medical Education Secretariat of the Faculty of Medicine and the Faculty of Engineering at UNAM were invited to participate. The invitation was extended one week prior to the start of the validation process, and the evaluation period lasted two weeks. Contact was made via instant messaging, during which the instrument's objective, its design process, the purpose of the validation, and the project's general context were explained, as well as the educational tool's objectives. This process aimed to obtain initial evidence of the instrument's content validity before its application in two pilot studies scheduled for October and November 2025.

2.6 Characteristics of the experts panel

A total of 13 experts were invited, with the following professional profiles: five doctors, four psychologists, three engineers and one educator. The number of judges was defined according to the methodological recommendations for content validity studies, which suggest a minimum of 6 to 10 experts to obtain stable estimates of the content validity index (18). Table 3 summarizes the characteristics of the expert judge panel.

Table 3. Characteristics of the expert judge panel (n = 13).

Variable	Category	n (%)
Years of experience in health sciences education	0–5 years	2 (15.4)
	6–10 years	4 (30.7)
	11–15 years	1 (7.7)
	16–20 years	2 (15.4)
	21–25 years	2 (15.4)
	>25 years	2 (15.4)
Appointment or role at UNAM	Full-time academic	3 (23.0)
	Academic of subject	1 (7.7)
	Academic official	1 (7.7)
	Part-time Academic	2 (15.4)
	Full-time academic and academic officer	2 (15.4)
	Part-time lecturer	2 (15.4)
	Full-time and part-time academic	1 (7.7)
	Lecturer and academic officer	1 (7.7)

2.7 Evaluation procedure

Each of the experts, blinded and independent, responded in a self-administered electronic form through the Google Forms platform, in which they evaluated the "Perception Questionnaire on the virtual patient Maria V". The judges assessed each of the 23 items to determine their relevance and suitability to the construct they were intended to measure, according to the

established criteria for content validity. To this end, each item was rated in terms of relevance using a four-point Likert scale: 1 = not relevant; 2 = somewhat relevant; 3 = relevant; 4 = very relevant. Additionally, the form included an open-ended section for experts to provide recommendations and qualitative comments aimed at improving the instrument. The open-ended questions were incorporated as complementary qualitative components of the instrument and not as items intended to generate a quantitative score. Their assessment by the judges was intended to estimate the relevance of their content to the construct of student perception of the AI-enhanced virtual patient, as well as their suitability for exploring aspects related to educational utility, opportunities for improving the tool, and its comparison with standardized human patients. Therefore, these items will not be considered for calculating scale scores, internal consistency, factor structure, or measurement models in subsequent phases.

2.8 Statistical analysis

The analysis of responses followed the methodology proposed by Polit et al. (18) and used Stata version 14 software. Content validity indices (CVI) were calculated at both the item level (I-CVI) and the overall scale level (S-CVI), using the arithmetic mean of the I-CVIs (S-CVI/Ave). The I-CVI was calculated by considering the proportion of judges who rated each item as relevant or highly relevant (Likert scale values 3 and 4). Following the recommendations of Polit et al. (18), items with I-CVI values ≥ 0.78 were considered to have adequate content validity. Additionally, the probability of inter-judge agreement was estimated, given that a limitation of the CVI is its inability to adjust for agreement attributable to chance. To correct for this bias, the modified kappa statistic (k*), proposed by Wynd et al., was calculated. (19) and taken up by Polit et al. (18), which adjusts the level of agreement considering the probability of coincidence by chance in the relevance category. The k* values were interpreted according to the criteria of Fleiss (20) and Cicchetti et al. (21), classifying them as acceptable, good, or excellent according to their magnitude. In accordance with Kane's validity framework (7), these analyses provide initial evidence to support the score inference by assessing the representativeness, relevance, and conceptual congruence of the items that make up the instrument.

2.9 Ethical considerations

This study was not submitted for review by an ethics committee because it was a methodological investigation of content validity focused on the evaluation of an instrument by expert judges, without direct educational intervention with students, without the collection of clinical, sensitive, or identifiable data from patients, and without posing any risks greater than minimal to the participants. The judges' participation was voluntary, self-administered, and aimed exclusively at assessing the relevance of the questionnaire items. Participants also signed a confidentiality agreement to safeguard the originality of the clinical scenario, protect the intellectual property of the material, and prevent premature bias in future applications of the instrument. Responses were analyzed in aggregate, without presenting individualized results. Review by expert judges helped strengthen the methodological rigor of the instrument and generate initial evidence of content-based validity before its application in pilot studies with student populations.

3. Results

Regarding the results of the content validity study, the 23 items of the "Perception Questionnaire about the Virtual Patient Maria V" were evaluated by 13 expert judges. The I-CVI values ranged from 0.69 to 1.00. Most items showed I-CVI values equal to or greater than 0.78, a criterion considered indicative of adequate content validity when six or more experts participate. In particular, several items reached an I-CVI of 1.00, reflecting unanimous agreement regarding their relevance. Item 14 showed an I-CVI value of 0.69, below the recommended cutoff point, indicating less agreement among the experts regarding its relevance. Similarly, item 9 showed an I-CVI of 0.77,

a value close to the minimum acceptable criterion; therefore, both items were considered subject to revision based on the qualitative observations made. Regarding the adjustment for probability of agreement attributable to chance, the modified kappa statistic values ranged from 0.66 to 1.00. According to the interpretation criteria proposed by Fleiss (20) and Cicchetti, most items presented kappa values classified as good to excellent (table 4).

Table 4. Content validity indices per item of the "Perception Questionnaire about the virtual patient María V."

Item	N	TO	I-CVI	PC	Modified Kappa (k*)
Item 1	13	11	0.85	0.01	0.85
Item 2	13	12	0.92	0.00	0.92
Item 3	13	13	1.00	0.00	1.00
Item 4	13	13	1.00	0.00	1.00
Item 5	13	13	1.00	0.00	1.00
Item 6	13	11	0.85	0.01	0.85
Item 7	13	13	1.00	0.00	1.00
Item 8	13	12	0.92	0.00	0.92
Item 9	13	10	0.77	0.04	0.76
Item 10	13	11	0.85	0.01	0.85
Item 11	13	13	1.00	0.00	1.00
Item 12	13	13	1.00	0.00	1.00
Item 13	13	11	0.85	0.01	0.85
Item 14	13	9	0.69	0.09	0.66
Item 15	13	12	0.92	0.00	0.92
Item 16	13	12	0.92	0.00	0.92
Item 17	13	12	0.92	0.00	0.92
Item 18	13	11	0.85	0.01	0.85
Item 19	13	12	0.92	0.00	0.92
Item 20	13	13	1.00	0.00	1.00
Item 21	13	12	0.92	0.00	0.92
Item 22	13	13	1.00	0.00	1.00
Item 23	13	12	0.92	0.00	0.92

Abbreviations: N = number of judges; A = number of agreements; I-CVI = content validity index per item; Pc = probability of agreement by chance; k* = modified kappa.

Items with kappa values below 0.75 coincided with those that presented lower I-CVI values. At the scale level, the content validity index calculated using the arithmetic mean of the I-CVIs (S-CVI/Ave) was 0.916, a value that exceeds the recommended criterion of 0.90 for considering adequate overall content validity. This result suggests that, overall, the instrument adequately represents the construct being assessed. Conversely, the content validity index calculated using the universal agreement method (S-CVI/UA) was 0.348. Most items remained unchanged as they presented adequate content validity values.

3.1 Qualitative analysis of the modified and eliminated items

As a complement to the quantitative indicators of content validity, the qualitative observations made by expert judges were analyzed to inform decisions regarding the retention, modification, or elimination of items. This analysis identified that some items required wording adjustments to improve their conceptual precision, while others had to be eliminated due to lower inter-judge agreement or because they did not fully correspond to the current functional characteristics of PV-IA.

Item 1 was modified because its initial wording referred to strengthening "clinical competencies for the OSCE," which could include broad components not directly assessed by the tool, such as physical examination, therapeutic decision-making, or procedural execution. Therefore, this wording was replaced with "clinical skills" to more precisely define the scope of training provided by the PV-IA, particularly regarding medical history taking, clinical reasoning, and communication during the interview.

Item 10 was adjusted to clarify the system's usability dimension. The original wording generally referred to the system's ease of use of the available tools; however, it was deemed necessary to specify the interaction elements more clearly. Therefore, the final version indicates that the system proved to be easy to use, allowing for simple interaction with the buttons and icons available for performing tasks. This adjustment made the operational component of the user experience more explicit.

Regarding the removed items, item 6 was withdrawn despite having obtained adequate I-CVI and modified kappa values. This decision was based on a criterion of conceptual and functional coherence, since the item's content attributed to the virtual patient a level of emotional response that did not fully correspond to the tool's current capabilities. In particular, it was considered that the item could lead students to assess an emotional or affective dimension that is not yet part of the actual functioning of the PV-IA. Therefore, its removal allowed for maintaining the correspondence between the instrument's items, the defined construct, and María V.'s current characteristics.

Item 9 was removed because it had an I-CVI of 0.77 and a modified kappa of 0.76, values close to the established minimum cutoff point. Although response speed can influence the interaction experience, qualitative observations suggested that the item could be interpreted in different ways, as it did not clearly differentiate between system latency, fluency in the conversation with the virtual patient, and the pedagogical continuity of the clinical simulation. For this reason, it was decided to remove it from the final version to avoid interpretive ambiguity.

Item 14 was eliminated due to its lowest content validity values, with an I-CVI of 0.69 and a modified kappa of 0.66. These results indicated less agreement among the judges regarding its relevance to the construct being assessed. Furthermore, qualitative observations suggested that the level of detail in the visual environment, while potentially influencing concentration and the perception of realism, was not a sufficiently central or specific indicator for assessing students' perceptions of the educational, clinical, and functional utility of PV-IA. Therefore, it was decided to remove it from the final instrument.

Item 16 underwent a wording adjustment to clarify the role of the avatar's gestures and expressions within the simulated clinical experience. The modification aimed to prevent the item from suggesting a direct clinical or diagnostic interpretation based on facial expressions and instead focus on the support these visual elements can provide for realism, the characterization of the virtual patient, and the understanding of the interaction.

Item 18 was also revised to clarify that tone of voice should be assessed based on its contribution to understanding the clinical information provided by the virtual patient, and not as an aesthetic or isolated characteristic of the system. This modification aimed to strengthen the connection between the auditory component of the interaction and the student's perception of the communicative and functional quality of the virtual patient interaction.

In total, four items were modified (1, 10, 16, and 18) and three were eliminated (6, 9, and 14), while the remaining 16 items were retained without substantial changes. As a result of this process,

the final instrument consisted of 20 items: 17 Likert-scale items and three open-ended questions. Table 5 summarizes only the items that were modified or eliminated after the quantitative and qualitative analysis.

Table 5. Items modified or removed after the content validation process.

Original item	Decision	Main reason	Quantitative justification	Qualitative/conceptual justification
1	Modified	Delimitation of the assessed competence	I-CVI = 0.85; k* = 0.85	"Clinical competencies" was replaced with "clinical skills" to avoid an overly broad reference to general clinical performance and to better define the training scope of PV-IA in the context of the OSCE.
6	Deleted	Conceptual incongruity with the current capabilities of PV-AI	I-CVI = 0.85; k* = 0.85	Although it obtained adequate scores, it was eliminated because it attributed to PV-IA a level of emotional response that does not fully correspond to its current capabilities.
9	Deleted	Value close to the cutoff point and interpretive ambiguity	I-CVI = 0.77; k* = 0.76	He did not clearly differentiate between technical latency, fluency in conversation, and pedagogical continuity of the simulation.
10	Modified	Precision of the system's ease of use	I-CVI = 0.85; k* = 0.85	The system's usability was improved through more specific wording regarding the simple interaction with the buttons and icons available for performing tasks.
14	Deleted	Low score on validity indices	I-CVI = 0.69; k* = 0.66	The level of detail in the visual environment was considered less central and specific to the construct being evaluated.
16	Modified	Conceptual precision of the role of gestures and expressions	I-CVI = 0.92; k* = 0.92	The wording was adjusted to refer to "the avatar's expressions," avoiding attributing an excessive diagnostic or emotional function to the virtual patient's gestures.
18	Modified	Accuracy of the auditory communicative component	I-CVI = 0.85; k* = 0.85	It was specified that the tone of voice should be assessed for its contribution to the understanding of clinical information and the communicative quality of the interaction.

Taken together, these results provide initial evidence of content-based validity, supporting the relevance and representativeness of the items as indicators of the student perception construct regarding PV-IA. This evidence contributes preliminarily to supporting the score inference within the validity argument.

4. Discussion

The results of this study provide favorable evidence of content-based validity for the "Perception Questionnaire on the Virtual Patient María V," at both the item and scale levels. Most items achieved I-CVI values equal to or higher than the recommended cutoff point, suggesting that experts considered the items relevant and representative of the construct related to student perception of the virtual patient (18). In particular, several items obtained I-CVI values of 1.00, reflecting unanimous agreement regarding their relevance, a criterion considered indicative of adequate content validity when six or more experts participate (13).

Furthermore, most items showed modified kappa values classified as good to excellent, indicating a high level of agreement among experts beyond chance. This finding strengthens the consistency of the assessments and supports the relevance of the items comprising the instrument. Items with kappa values below 0.75 coincided with those showing lower I-CVI values, confirming the consistency between both indicators and reinforcing the methodological decision to revise or eliminate these items. From the perspective of validity, this consistency helps support the score inference, demonstrating that the evaluation process was sensitive in discriminating between more and less relevant items (5-6).

The refinement of the instrument highlighted the need to integrate quantitative, qualitative, and conceptual criteria during the content validity processes. Items 9 and 14 were eliminated primarily due to their lower I-CVI and modified kappa values, as well as qualitative observations from the judges related to interpretive ambiguity or lower centrality to the construct. In contrast, the elimination of item 6 was not due to low psychometric performance, but rather to a criterion of functional congruence, given that it referred to the virtual patient's emotional response, a characteristic that María V. has not yet consistently reproduced. Retaining this item would have required students to assess a dimension insufficiently represented in the interaction experience, which could affect the interpretation of the scores. Taken together, these decisions show that the retention of an item should not depend solely on statistical cut-off points, but also on its conceptual coherence, its suitability to the instrument's educational purpose, and its correspondence with the actual capabilities of the technology being evaluated.

At the scale level, the S-CVI/Ave score exceeded the recommended criterion of 0.90, indicating adequate overall representativeness of the construct being evaluated and providing sufficient evidence of content validity (18). In contrast, the index calculated using the universal agreement method (S-CVI/UA) was lower, at 0.348. This difference is methodologically relevant, since the S-CVI/Ave reflects the average level of agreement among experts regarding the relevance of the items, while the S-CVI/UA requires absolute consensus on each item, thus constituting a considerably stricter criterion (18). In instruments with a large number of items and broad panels of judges, as in the present study, the probability of achieving universal agreement decreases substantially, even when individual items are relevant. Therefore, the low S-CVI/UA score should not be interpreted as evidence of insufficient content validity, but rather as a reflection of the restrictive nature of this indicator. Additionally, it allowed the identification of items with less consensus among experts and guided their review, modification or elimination during the instrument refinement process.

The main contribution of this study is that it proposes and documents an explicit methodological process for constructing and refining an instrument designed to evaluate student perceptions of an AI-based virtual reality simulation (VV-AI). Furthermore, this validity approach is relevant because the rapid integration of AI-based tools in medical education demands not only an assessment of their technical feasibility but also the development of instruments that allow for a rigorous interpretation of students' learning experiences. In this regard, the questionnaire

developed does not simply measure technological acceptance but also incorporates educational and clinical dimensions inherent to simulation, such as the quality of interaction, perceived realism, and the educational value of the experience.

In comparison with some international studies related to virtual patients in the OSCE, in which explicit content-based validity procedures are not always described, this study transparently describes the process of constructing, peer reviewing, and refining the instrument. As documented in the literature review conducted for this research, several previously used instruments do not report systematic content validation processes by expert judges equivalent to that developed in this study. In this sense, the main contribution of the present research lies in the transparency, systematization, and methodological rigor applied during the validation process (22-24).

The rigorous methodology implemented allowed not only the development of a functional tool to assess student perceptions of AI-mediated PV, but also the generation of a replicable methodological roadmap for future research (25). This aspect represents a relevant contribution to medical education, by establishing a benchmark for the design and validation of metrics aimed at evaluating AI-mediated educational experiences.

From an educational perspective, the incorporation of AI-powered simulations transforms the paradigm of simulated learning, allowing students to practice anamnesis, clinical reasoning, and communication skills in safe environments with high dialogic fidelity (26). However, the pedagogical effectiveness of these tools depends largely on student acceptance, perceived usefulness, ease of use, and the realism they experience, as well as on the alignment of the construct being measured. In this context, having instruments with initial evidence of content validity allows for a more systematic evaluation of the interaction between students and these technologies, contributing to optimizing clinical scenarios, identifying potential risks, such as inaccurate responses generated by artificial intelligence, and strengthening feedback and learning processes.

On the other hand, from a curricular perspective, the content validity findings allow us to identify dimensions with the potential to guide the gradual incorporation of virtual patients into clinical training programs. The items with the greatest consensus among experts were related to educational value, the quality of clinical interaction, the clarity of verbal communication, clinical reasoning, and overall satisfaction—dimensions that can be linked to OSCE training activities, medical history taking, doctor-patient communication, and preparation for simulated clinical scenarios. In this sense, the instrument not only allows us to assess the acceptance of a technological tool but also to generate useful information for curriculum design, the improvement of clinical simulation experiences, and decision-making regarding the progressive integration of virtual patients into medical education.

Limitations of the study

Content validity is only one of the sources of validity evidence considered in contemporary evaluation frameworks (27). In this sense, the reported results do not yet allow us to establish evidence related to the internal structure of the instrument, its relationships with other variables, or criterion validity. Although the panel of experts was made up of professionals with diverse experience in education and health sciences, the interpretation of the results continues to depend on the subjective judgment of the evaluators. Also, the study focused exclusively on evaluating student perceptions of a PV-IA within a specific context, so generalizing the results to other educational or technological environments should be approached with caution. Another limitation of this study is that the quantitative assessment by the judges focused on the relevance and representativeness of the instrument's items. While the qualitative feedback identified opportunities for improvement related to the clarity, coherence, and conceptual consistency of the items, these dimensions were not

assessed using independent scales that would formally quantify the degree of agreement among the experts. Consequently, the evidence obtained is primarily focused on content validity associated with the relevance of the items. Future research could strengthen this process by incorporating differentiated assessments of clarity, coherence, and sufficiency. Furthermore, the proposed dimensions should be interpreted as theoretical content domains and not as empirically confirmed latent factors. Consequently, it is not yet recommended to interpret scores by subscale or to assume a factorial structure of the instrument. To strengthen the validity argument, subsequent studies should evaluate the internal structure, internal consistency, behavior of the dimensions, and their association with other variables by applying the instrument to a student population (5-6).

5. Conclusions

- This work provides an initial methodological framework for developing instruments to evaluate AI-mediated educational experiences in clinical simulation. Its main contribution lies in integrating a technologically accepted theoretical framework with dimensions specific to simulation-based medical education, as well as documenting a systematic process of expert review and item refinement. Future research should apply the instrument to student populations to provide new sources of validity evidence, including internal structure, internal consistency, and relationships with other variables.
- The removal of items 9, 14, and 6 reflects the importance of considering not only statistical criteria but also the conceptual and functional coherence of the instrument during the validation process. Overall, this process allowed for the refinement of the questionnaire and strengthened its suitability for subsequent pilot studies. The findings of this study provide initial evidence of content-based validity for the "Perception Questionnaire on the Virtual Patient María V," contributing to the development of methodologically sound tools for evaluating student perceptions of emerging technologies in medical education.

Funding: There has been no funding.

Declaration of conflict of interest: The authors declare that they have no conflict of interest.

Authors' contributions: Conceptualization: CASA, AOM, ACSV, JATM, SCC, and FGL. Methodology: CASA and FGL. Software: SCC and MSR. Validation: AOM, AC, JATM, AIMP, SCC, and FGL. Formal analysis: CASA, AC, and AOM. Research: CASA, MSR, and FGL. Resources: ACSV, SCC, and MSR. Data curation: CASA. Writing - original draft: CASA, AOM, AC, JATM, AIMP, SCC, MSR, and FGL. Writing - revision and editing: CASA, AOM, AC, ACSV, JATM, AIMP, SCC, MSR, and FGL. Visualization: AOM, AC, and AIMP. Supervision: AOM and ACSV. Project management: CASA, AOM, ACSV, and SCC. Fundraising: not applicable.

6. References.

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 interterrater 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