BMC Medical Education

RESEARCH Open Access

Development and validation of virtual patients for assessing history taking and communication skills among medical students

Nurul Ain Mohd Salim[1], Azreen Azman[2], Puteri Suhaiza Sulaiman[2], Nurhanis Syazni Roslan[3], Rafidah Hod[1] and Siti Khadijah Adam[1*]

Abstract

Background The integration of virtual patients (VPs) into medical education is expanding globally, enhancing traditional training and supporting the development of clinical skills. In Malaysia, adoption is progressing, although dependence on imported platforms may compromise cultural and linguistic authenticity. We developed three culturally adapted VPs reflecting the local healthcare context. The VP application, featuring a custom virtual reality (VR) environment, 3-dimensional (3D) avatars and speech functions, was deployed on the Oculus Meta Quest 2 headset. This device provides an immersive, user-friendly experience, allowing students to interact with VPs via hand-held controllers.

Methods The developed VPs were assessed in terms of their content, face, construct, and concurrent validity, as well as the reliability for use in history-taking and communication skills assessment. Data analysis was performed using IBM SPSS Statistics software version 25.

Results The VPs showed acceptable content and face validity, indicating that the case design, interface, and simulated environment were appropriate and realistic. However, reliability and other validity measures, including construct validity and concurrent validity, were found to be limited.

Conclusion These findings may have been affected by factors such as small sample size, variability among students and examiners, limited scenario complexity, and technical issues related to speech recognition and system responsiveness. Addressing these limitations in future research could improve the generalizability of the results, while refinements in scenario design, artificial intelligence (AI) integration, and user interfaces are key to maximizing the pedagogical potential of VPs.

Keywords Virtual patients, Medical education, Virtual reality, History taking, Communication skills

*Correspondence:
Siti Khadijah Adam
sk.adam@upm.edu.my
[1]Department of Medical Education, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[2]Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[3]Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia, Kota Bharu, Kelantan, Malaysia

Introduction

Strong communication skills are essential for medical students, as they play a pivotal role in academic achievement, interpersonal interactions, and adaptation to professional settings [1]. Effective communication with patients, whether by medical students or healthcare professionals, is fundamental for fostering meaningful interactions and gathering accurate clinical information. To support the development and assessment of these skills, simulated patients are commonly employed, particularly prior to students' exposure to real clinical encounters.

Simulated or standardized patient (SP) refers to an individual trained to portray a patient with a specific condition for educational purposes. The use of SPs in medical education was first introduced in the 1960s as a standardized approach for assessing clinical competence [2]. SPs offer medical students a valuable opportunity to develop communication skills, enhance clinical reasoning, and engage with a variety of patient scenarios within a controlled and safe environment [3]. These individuals are specifically trained to consistently replicate particular medical cases, allowing learners to practice repeatedly without posing any risk to actual patients. The standardized nature of SP scenarios ensures equitable learning experiences across student cohorts, promoting uniformity in training and assessment. Moreover, SPs help bridge the gap between theoretical instruction and practical application, contributing to a more authentic learning experience. By replicating real-life clinical situations, SPs help build student confidence in both procedural tasks and decision-making processes, better preparing them for the complexities of actual patient care [4, 5].

The transformation of SPs into VPs has evolved alongside technological advancements and innovations in medical education. The integration of VPs as emerging digital tools made it possible to simulate realistic clinical encounters within safe and controlled environments. VPs are computer-based simulations of clinical scenarios designed to support the teaching and assessment of medical competencies. These simulations differ in levels of fidelity, interactivity, and realism, from simple, text-based formats to complex, AI-enhanced environments. Low-fidelity VPs typically rely on text-based cases, decision-tree models, or rule-based systems, which often have limited interactivity. They are frequently utilized for foundational learning, such as case-based learning and early-stage clinical reasoning [6]. On the other hand, high-fidelity VPs incorporate advanced technologies such as AI, natural language processing (NLP), speech recognition and physiological modeling, offering a more dynamic and realistic learning experience. An example is the Oxford Medical Simulation used in Hull York Medical School, United Kingdom. It is a VR-based package designed for head-mounted displays (HMDs) with auditory input. It offers immersive scenarios where students can examine, investigate, and treat VPs in real time, with dynamic physiological responses to their actions [7].

The use of VPs in medical education has seen a global rise, with institutions tailoring these tools to suit specific educational needs. As highlighted by Plackett et al. (2022), VPs are increasingly employed to mitigate the limitations posed by the scarcity of SPs [8]. Medical schools worldwide have successfully incorporated VPs into their medical curricula to supplement traditional training, enabling students to develop clinical competencies in structured and replicable scenarios [9, 10]. Additionally, VPs are used to facilitate cultural competence training [11, 12]. In Malaysia, the integration of VPs is gradually advancing as medical institutions adopt existing platforms from abroad to enrich their educational programs. However, reliance on imported VPs may limit the authenticity of patient interactions due to a lack of alignment with local cultural and linguistic contexts. This underscores the need for locally developed VPs that reflect Malaysian societal norms, communication styles, and healthcare practices. In response, we designed culturally adapted VPs tailored to the Malaysian healthcare landscape, enhancing the effectiveness of training in delivering contextually relevant and culturally sensitive patient care.

Development of VPs

The design of our VPs was informed by the key components of communication skills crucial for effective doctor–patient interactions [13]. Cultural and linguistic elements were carefully integrated to ensure alignment with the Malaysian context and to enhance the relevance of the VPs within local medical education curricula. These contextualized inputs played a key role in shaping the development of the 3D avatars and the associated speech functionalities in the VPs application. The development was carried out by a multidisciplinary team comprising medical lecturers with extensive experience in teaching and assessing history-taking, medical educationists with expertise in simulation-based learning, and technical experts in computer graphics, VR, speech recognition, and text mining. This combination of educational and technical expertise ensured that the VP system was both clinically authentic and technologically sound.

Three male VPs were developed, each representing a distinct medical scenario and reflecting the major ethnic groups in Malaysia, namely Malay, Chinese, and Indian, to mirror the cultural diversity commonly encountered in local clinical settings. Beyond visual representation, each VP incorporated culturally embedded communication patterns and behavioral nuances consistent with the respective background. These adaptations were expressed through variations in dialogue structure, tone, language expressions, and non-verbal cues, while ensuring that all scenarios maintained comparable levels of complexity and clinical reasoning demands. This culturally contextualized design aimed to provide learners with authentic and inclusive patient interactions representative of Malaysia's multicultural healthcare environment.

Each VP was assigned a specific clinical scenario, i.e., chest pain, urinary tract infection and gastritis, reflecting commonly taught conditions in undergraduate medical training. For each case, dialog scripts were crafted to include plausible student inquiries and corresponding VP responses, enabling realistic and pedagogically sound interactions. Unfortunately, the development of a Malay-language version of the VPs was not feasible within this study due to the absence of a robust speech-recognition engine capable of accurately processing the Malay language. Therefore, all VPs interactions were conducted in English, which is widely used as a second language in Malaysia and serves as the primary medium of instruction in Malaysian medical education. This technical limitation constrained the inclusion of a fully localized linguistic experience. However, the use of English remained pedagogically appropriate and consistent with the instructional and assessment practices in Malaysian medical schools.

The Unity platform, which is widely used in computer science and game development, offers a robust framework for creating interactive 2D and 3D digital environments. It equips developers with versatile tools to build video games, simulations, and applications for VR and augmented reality (AR) [14]. This technology enables the generation of realistic patient avatars, allowing users to engage with them through the VPs interface. In this study, a virtual consultation setting was developed using high-fidelity graphics designed to simulate a realistic clinical environment. The layout included an L-shaped seating arrangement for the doctor and VP, a consultation desk, and other relevant furnishings, all strategically placed to reflect appropriate interpersonal distances and to promote a sense of privacy during patient interactions.

A Syn Bot, or Synthetic Bot, is a chatbot engineered to simulate human-like conversations [15]. These bots leverage AI and NLP to interpret user input and generate meaningful responses. In this study, the Syn Bot was configured to receive speech input and match it against predefined dialog scripts. Prior to enabling voice interaction, the scripted dialogs were converted into a structured format compatible with the Syn Bot Framework. This framework serves as a conversational engine that is responsible for interpreting user intent, processing input, and generating suitable responses to replicate realistic patient interactions. It functions similarly to a central processing unit, guiding the VP's responses on the basis of user prompts [16]. The Syn Bot framework uses both NLP and natural language understanding (NLU) to accurately recognize and interpret spoken or written inputs. NLP, a branch of AI, allows the system to analyze human language, whereas NLU focuses on understanding the intent and contextual meaning behind the user's statements [17]. In our project, we integrated Wit.ai, an NLP tool developed by Meta, to enable the VP system to comprehend user interactions. The overarching intelligence of the system is governed by AI, which facilitates dynamic and human-like conversational exchanges [18].

Upon receiving user input, the chatbot analyzes it to determine the underlying intent or purpose of the message. For example, when a user asks, "*Why are you here today?*", the system may interpret the intent as identifying the patient's 'chief complaint'. Once the intent is recognized, the chatbot generates an appropriate response by considering the ongoing conversational context. At this stage, dialog management plays a critical role. It involves tracking the flow of the interaction, preserving contextual continuity, and selecting suitable responses or actions for the VPs to deliver. The VPs application maintains accuracy in matching user questions with appropriate responses by referencing a set of predefined intents, ensuring coherence and realism in the simulated clinical interaction.

Each 3D avatar was enhanced with articulated skeletal structures and dynamic body animations, complemented by culturally authentic voiceovers to emulate realistic patient behaviors, gestures, and accents. The underlying AI algorithms empowered the VPs to select appropriate facial expressions, gestures, and verbal responses autonomously during user interactions. Their replies were programmed via simple, user-friendly terminology, and they were designed to listen and respond contextually to user questions. Nonverbal communication was also integrated, including eye contact, facial expressions of pain, body language, and hand gestures, to indicate the location and radiation of pain. The VPs could additionally convey understanding or disagreement through head nods or shakes. Figure 1 illustrates the designed 3D virtual consultation room and one of the VP avatars.

The VP application, including the custom-designed VR environment, 3D avatars, and integrated speech functionalities, was deployed on the Oculus Meta Quest 2, a VR headset developed by Meta. This device offers an immersive and user-friendly VR experience, known for its intuitive control, interactivity, and adaptability across various applications. Users can effortlessly launch the application and engage with virtual patients via the headset's hand-held controllers.

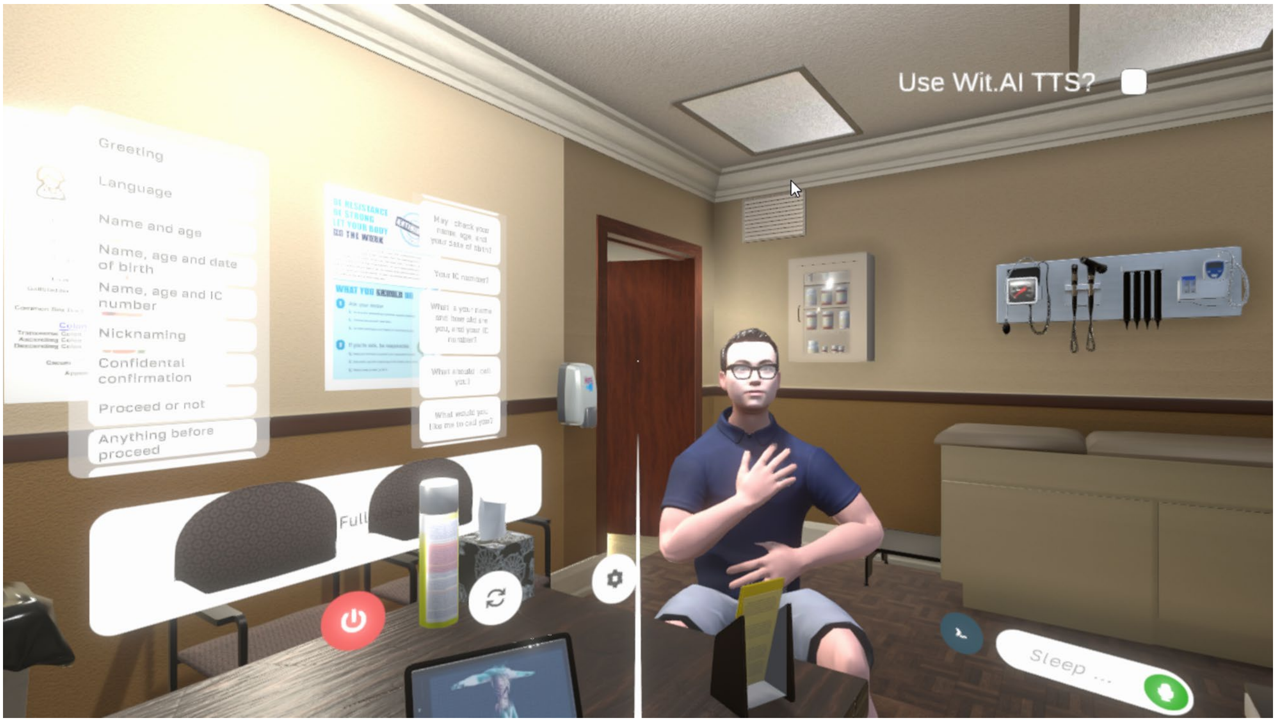

Fig. 1 Virtual patient avatar in the designed 3D consultation environment

Methods

The developed VPs were then evaluated for their content, face, construct, and concurrent validity tests as well as their reliability for assessing medical students' history taking and communication skills. The study was performed at the Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Malaysia, and involved undergraduate medical students and lecturers teaching the programme.

All participants were informed about the objectives, procedures, potential risks, and benefits of the study prior to participation. Written informed consent was obtained from each participant before data collection. Participation was voluntary, and participants were informed of their right to withdraw from the study at any stage without any academic or personal consequences.

Content validity

Content validity refers to the extent to which an instrument accurately measures the content it is designed to assess [19]. In this study, it evaluates whether VPs, particularly medical scenarios and responses by VPs, are relevant to the subject matter. A set of questionnaires related to the content and responses of the VPs were derived and modified from previous studies [20–23].

The modified questionnaire included 15 items, which the experts, who are medical lecturers, were required to rate. The items were rated via a Likert scale, with scores of 1 indicating that the content was not relevant, 2 indicating that it was somewhat relevant, 3 indicating that it was relevant, and 4 indicating that it was very relevant. The questionnaire comprised items that evaluated the relevance, appropriateness and educational usefulness of the VPs. Specifically, the experts assessed the VPs' feedback, language and tone, verbal, paraverbal and nonverbal communication aspects of medical history taking. The questionnaire also examined the suitability of each case for evaluating students' questioning ability, realism of patient responses and overall usefulness of the VPs in teaching and assessing communication skills in undergraduate medical training. They also rated the VPs ability to simulate a realistic scenario that allowed the student to take a full and appropriate patient history relevant to the specific medical problem, ensuring all essential patient data could be elicited. This approach provides a quantitative assessment of the lecturers' perceptions with respect to the suitability and comprehensiveness of the content and the appropriateness of the responses by the VPs.

Prior to evaluation, all experts were provided with a brief description of the study objectives and the constructs and domain used in the questionnaire to ensure consistency in interpretation. They were introduced to the VPs and instructions on their usage. They interacted with each of the three VPs sequentially and subsequently completed the questionnaire. They were also invited to provide open-ended feedback on the overall appropriateness and relevance of the VPs, including the cultural realism and suitability of each VP's background, communication style and behavioral nuances, as well as any other suggestions for improvement. The feedback gathered was subsequently used to refine the dialogues and non-verbal expressions, ensuring that each VP and their responses were relevant, contextually appropriate and culturally sensitive, accurately reflecting patient portrayals commonly encountered in Malaysian clinical settings.

To measure content validity, the content validity index (CVI) was calculated. The CVI indicates the level of agreement in relation to the validity of each item and the scale, with values ranging from zero to one. The acceptable CVI value is at least 0.78 [24, 25]. The assessment occurs both at the item-level content validity index (I-CVI) and the scale-level content validity index (S-CVI/Ave). The formulas for both the I-CVI and S-CVI/Ave are as follows:

$$I-CVI = \textit{the number of individuals who agreed that an item was relevant } [\textit{rating } 3 \textit{ or } 4] / \textit{total number of experts}$$

and

$$S-CVI/Ave = \textit{average of the } I-CVI \textit{ scores/the number of items}$$

The CVI calculation was conducted via Microsoft Excel. Refinements to the content of the scenario cases and the responses from the VPs were derived on the basis of feedback from the experts.

Face validity

Face validity refers to the extent to which a simulator appears realistic and appropriate in representing the actual task it is intended to replicate [26]. This study involves assessing the appearance of VPs, including the satisfaction level of users, particularly medical students, and identifying areas for potential enhancement. A set of questionnaires related to the realism of users' experiences with VPs was derived and modified from previous studies [20–22, 27]. Undergraduate medical students were recruited to gain experience in handling and communicating with VPs. They were then instructed to assign ratings to each item in the questionnaire on the basis of the VPs' clarity and comprehensibility.

The questionnaire comprised 17 items that evaluated the realism, usability and clarity of the VPs. Specifically, the students assessed the visual quality and realism of the 3D environment, the smoothness of movement, the clarity and responsiveness of voice-based interactions and the quality of verbal and paraverbal expressions demonstrated by the VPs. The questionnaire also examined the simplicity, ease of use, and user-friendliness of the VP interface, as well as the overall suitability of the VPs for assessing communication skills among medical students. The items were rated via a Likert scale, with scores of 1 indicating that they were not clear and incomprehensible, 2 indicating that they were somewhat clear and comprehensible, 3 indicating that they were clear and comprehensible, and 4 indicating that they were very clear and comprehensible.

Prior to evaluation, all students were provided with a brief description of the study objectives and the constructs and domain used in the questionnaire to ensure consistency in interpretation. They were introduced to the VPs and instructions on their usage. They interacted with the VPs and subsequently completed the questionnaire. They were also invited to provide open-ended feedback on the overall realism and clarity of the VPs, including the usability and any other suggestions for improvement for further refinement.

To measure face validity, the face validity index (FVI) was calculated. The FVI was established with a minimum threshold of 0.80 for acceptability [28]. The FVI encompasses both the item-level face validity index (I-FVI) and the scale-level face validity index (S-FVI/Ave), which are designed to assess the clarity and comprehensibility of the items. The formulas are as follows:

$$I-FVI = \textit{the number of individuals who agreed that an item was clear and comprehensible } [\textit{rating } 3 \textit{ or } 4] / \textit{total number of students}$$

and

$$S-FVI/Ave = \textit{average of the } I-FVI \textit{ scores/the number of items}$$

The FVI calculation was conducted via Microsoft Excel. Improvements to the features of the VP application were made on the basis of the students' feedback.

Reliability

Reliability refers to the consistency and stability of a measurement, denoting that when good reliability is achieved, users obtain the same results when they repeat the assessment [29]. In this study, the test aimed to assess the reliability of VPs by evaluating their consistency across various scenarios and the performance of medical students.

An objective structured clinical examination (OSCE) comprising three stations was conducted with first-year medical students. Each station involved each VP, which lasted for six minutes. The students were assessed while being evaluated by examiners through a laptop stream from the Oculus Meta Quest 2 device. OSCE checklists were provided to assess the students' performance at each station, with a focus on their history-taking and communication skills. These checklists were developed in accordance with the format used at UPM. Following the examination, the reliability of the OSCE scores for all the students was determined via Cronbach's alpha value.

Construct validity

Construct validity can be demonstrated when a test measures the construct it claims to be measuring. Specifically, one group possessing the construct performs better than another group lacking the construct [30]. In this study, the extent to which VPs are able to differentiate between different levels of expertise is investigated.

An OSCE comprising three stations was conducted, similar to the reliability test. In this study, two different groups of students were assessed, namely, first-year and third-year undergraduate medical students. The same VPs and OSCE checklists were used in both groups. The performances of first-year and third-year students were compared. It was expected that the performance of third-year students, who are more experienced in taking history and communicating with patients, would be better than that of first-year students.

Concurrent validity

Concurrent validity refers to the estimation of individual performance on different assessments, where one assessment is new and the other is well established and proven to be valid [31]. In this study, the researcher aimed to assess the performance of medical students via two distinct methods: VPs, a novel approach, and human SPs, a traditional method in OSCE examinations.

Two different OSCEs comprising three stations were conducted with the same group of students. One examination involved VPs, whereas another examination utilized human SPs. The same scenario and OSCE checklists were used in both examinations. The correlation of student performance between VPs and human SPs was determined. It was expected that the students' performance would correlate regardless of the type of patients involved in the OSCE.

Data analysis

All the data were analyzed via IBM SPSS Statistics software version 25. The distribution of the data for construct and concurrent validities was first analyzed via the Kolmogorov–Smirnov test. Normally distributed data was analyzed via a parametric test and are presented as the mean ± standard deviation. Any p-values of < 0.05 were considered significant.

Results

Content validity

A total of 12 lecturers participated as experts in the test. The panel was well balanced in terms of gender and seniority, ranging from 3 to 15 years of work experience and representing a diverse range of academic and clinical backgrounds. Experts were drawn from both pre-clinical and clinical disciplines, such as Anatomy, Physiology, Obstetrics and Gynecology and Family Medicine. All experts were qualified medical doctors with substantial experience in teaching and assessing history-taking skills, as well as in conducting patient interviews in real clinical settings. This multidisciplinary composition ensured that the content validity evaluation encompassed both educational and clinical perspectives relevant to the intended learning context.

The I-CVI scores for all 15 items in the questionnaire varied between 0.83 and 1. The S-CVI/Ave value, which is the average of the I-CVI scores, was found to be 0.93. These values exceed the 0.78 threshold [24, 25], indicating that the content within the VPs is deemed acceptable. Furthermore, these findings suggest that the responses provided by the VPs are appropriate for evaluating students' history-taking and communication skills.

Face validity

There were 43 medical students who participated as raters in the test, which provided a sufficiently large group to gather diverse and meaningful feedback on the VPs. They represented a well-balanced distribution of gender and ethnicity, reflecting the diversity of the medical student population. All students had at least one year of experience in learning and practicing history taking skills, including prior exposure to SP encounters as part of their training. This background ensured that the students were adequately familiar with conventional history taking methods, allowing them to provide informed and meaningful feedback on the realism and usability of the VP interactions.

The I-FVI scores for all 17 items in the questionnaire ranged from 0.86 to 1. The S-FVI/Ave, which represents the average of the I-FVI scores, achieved a value of 0.96. These scores exceed the 0.80 threshold [28], indicating that the features within the VPs are clear and comprehensible. This shows that the VPs were perceived as realistic and appropriate for student interaction, suggesting good usability and acceptance among learners.

Reliability

A total of 42 students participated in the OSCE to measure the reliability of the VPs. Table 1 shows the mean score of each OSCE station. Cronbach's alpha value is 0.415, which indicates that VPs have low reliability [32].

Table 1 Mean scores for each OSCE station utilizing VPs and the cronbach's alpha value ($n = 42$)

OSCE station	Mean score (SD)	Cronbach's alpha
Station 1 (chest pain)	81.71 (9.67)	0.415
Station 2 (urinary tract infection)	83.73 (9.08)	
Station 3 (gastritis)	78.94 (9.51)	

SD standard deviation

Table 2 Comparison of OSCE performance between first-year and third-year medical students utilizing VPs

Group	*n*	Mean score (SD)	df	*p* value
First year	42	80.74 (6.52)	62	0.219
Third year	22	82.83 (6.08)		

n frequency, *SD* standard deviation, *df* degrees of freedom

Table 3 Mean OSCE performance score when actual SPs and VPs are utilized ($n = 42$)

Type of patients	Mean score (SD)
Actual SPs	83.66 ± 8.29
VPs	80.74 ± 6.52

n frequency, *SD* standard deviation

Construct validity

A total of 42 first-year medical students and 22 third-year medical students participated in the OSCE. Table 2 shows the mean OSCE scores across all the stations for each group. Independent t-tests revealed no significant difference ($p = 0.219$), although third-year students presented slightly better results than first-year students did. This suggests that third-year students, who were exposed to real patients as well as SPs and opportunities to refine their history-taking skills, performed better than first-year students did, but the difference was not statistically significant.

Concurrent validity

Table 3 shows the mean scores of the OSCE across all stations utilizing actual SPs and VPs. Pearson correlation coefficient analysis revealed that $r = 0.094$ ($p = 0.552$). This suggests that while there is a minor positive relationship between the performances with SPs and VPs, the correlation is not substantial enough to be considered significant.

Discussion

Our findings indicate that the content and face validity scores for the VPs were high, reflecting the appropriateness and relevance of both the case content and the VPs' responses. Additionally, the design features of the VPs and the virtual consultation room were perceived to closely resemble real-life clinical scenarios. These results suggest that VPs hold promise for further application in educational settings, including structured assessments such as OSCEs.

Unfortunately, the reliability scores of the VPs across all three OSCE stations were low. Several factors may have contributed to this, including the small number of participating students and the limited number of VP scenarios included in the assessment. Previous reports suggest that increasing the number of stations or VP cases may improve reliability. For example, Yudkowsky et al. (2019) emphasized that reliability can be enhanced by increasing the number of assessment points [33]. In the present study, only three VP cases were developed, primarily because of the significant time and resources needed. Creating effective VPs involves complex programming to process user speech input accurately, which necessitates training machine learning models to recognize patterns and generate appropriate responses [34–36]. Moreover, substantial budgeting is required for software development and avatar modeling. Prior studies have emphasized that adequate financial investment at the early stages is essential for producing high-fidelity, educationally relevant VR simulations [37, 38].

The construct validity results showed that third-year students performed better on average than first-year students did, although the difference was not statistically significant. This may be due to the small sample size, which reduces the statistical power and increases the risk of overlooking true differences [39]. A larger sample may help confirm whether the observed trend reflects meaningful performance differences on the basis of clinical experience. Differences in clinical exposure and familiarity with the VP platform may also have influenced outcomes. As Botezatu et al. (2010) noted, variation in simulation performance often stems from unequal familiarity with the tool [40]. Additionally, if VP scenarios are not well aligned with students' training levels, their ability to differentiate competence may be limited [41]. External factors such as stress, time constraints, or technical issues may also influence their performance.

Our findings also demonstrated that although students may perform well in history-taking with SPs, this does not necessarily translate to better performance with VPs during assessments. This may be due to limited training and familiarity with the VP platform, which could have affected students' ability to engage effectively with the system. Previous studies support this observation. Fink et al. (2023) reported that students perform better with SPs than with VPs, likely because of the additional help provided by the SPs [42]. The less favorable results in this study may be influenced by examiner variability and the technical limitations of the VP system. Although examiner training was provided and a structured instrument was used, the inherent subjectivity in evaluating complex skills like medical communication remains a persistent challenge. While VPs are designed to simulate clinical encounters, they may not yet capture the full complexity, responsiveness, and interactive depth of real-time human SPs. This highlights the ongoing challenge of developing virtual tools that authentically replicate clinical realism.

The results may also be influenced by the characteristics or experiences of the students themselves. Several studies have noted that students often perceive VPs as less realistic due to the absence of body language, facial expressions and emotional responsiveness, which can lead to discomfort and lower performance [43–45]. While students found the VPs realistic and comprehensible based on face validity, the system may not have captured the nuanced interpersonal dynamics, emotional responsiveness and adaptive dialogue that typically occur in human SP encounters. Furthermore, Maicher et al. (2019) emphasized that accurate speech recognition remains critical; students who are nonnative English speakers or have learning difficulties may face additional challenges when interacting with VPs, affecting their ability to adapt to this assessment method [46].

Although certain psychometric indicators, particularly reliability, were less robust, the findings collectively highlight the educational potential of VPs. The strong evidence for content and face validity demonstrates their authenticity and relevance as simulated clinical encounters. While the current version may not yet be suitable for use as an independent summative assessment tool, VPs can play a valuable role in formative learning, enabling students to practice clinical reasoning, communication and history-taking skills in interactive and feedback-rich simulations. Integrating VPs alongside traditional SPs and real patient sessions can enrich students' experiential learning and support progressive development of clinical competence.

Differences in participants' backgrounds may have influenced the validation outcomes in various ways. For instance, experts from different disciplines and with varying levels of teaching experience may emphasize different aspects of realism, clinical accuracy or educational value when evaluating VPs. Likewise, students' cultural and linguistic backgrounds could affect their comfort level and communication style during VP interactions. The incorporation of cultural and linguistic elements in the VP design aimed to enhance authenticity and contextual relevance. Although these potential influences were not examined in the present study, acknowledging their presence is important for interpreting the findings and guiding future research. Subsequent studies should explore how user demographics and cultural–linguistic diversity influence perceptions, engagement, and performance outcomes in VP–based learning and assessment.

VPs and simulations, in general, provide a safe and controlled environment for students to practice clinical skills without posing any risk to real patients [47]. Simulation-based training has been shown to enhance clinical competence and decision-making by enabling repeated practice in standardized conditions, even for complex scenarios [48]. VPs help reinforce learning and improve diagnostic accuracy, effectively bridging the gap between theoretical knowledge and clinical application [8, 42, 47].

This study represents one of the earliest efforts to develop and validate a fully immersive, AI-driven VP system within a Malaysian medical education context. Unlike many existing VP platforms that rely on text-based or screen-mediated interactions, our system integrates voice-enabled communication through the Syn Bot framework, powered by AI, NLP and NLU technologies. The incorporation of Wit.ai, a speech-recognition tool, enables the VPs to interpret user intent and respond dynamically, thereby mirroring authentic clinical dialogue and enhancing interactional realism. Collectively, this integration of immersive VR technology and intelligent conversational capability advances the design of VPs beyond conventional 2D or scripted models. The immersive 3D consultation environment strengthens learner engagement and sense of presence. These unique features distinguish the present work from previous studies conducted primarily in Western contexts and underscore its potential relevance for broader adaptation in similar educational environments.

A major challenge encountered in this study was technical, particularly involving limitations in speech recognition, gender representation, and language compatibility. Issues with speech recognition and text-to-speech functionality can disrupt history-taking by generating irrelevant or inaccurate responses, often caused by limited input flexibility and user pronunciation variations. These technical challenges were identified through observations during student–VP interactions. In addition, technical aspects such as the usability and comfort of HMDs significantly influence the user experience in VR-based VP interactions.

This study was also limited to a single medical school, which restricts the generalizability of the findings. Differences in institutional assessment practices and learner demographics may influence outcomes. Furthermore, the small number of student participants and VP scenarios was a result of time and budget constraints.

Future research should aim to address several limitations identified in this study. Increasing the sample size and incorporating students from various universities would improve the generalisability of the findings and provide more rigorous statistical analysis. Increasing the quantity and diversity of VP scenarios, especially those exhibiting varied degrees of complexity, could enhance the tool's reliability and ability to assess different facets of clinical competence. Integrating a wider array of clinical disorders and different difficulty levels may enhance the assessment's discriminatory efficacy. Future development efforts should focus on enhancing the AI and NLP competencies of VPs. Improving voice recognition accuracy and emotional responsiveness will facilitate a more authentic and immersive simulation experience, hence narrowing the divide between virtual and actual patient interactions.

Conclusion

This study presents the development and initial validation of VPs designed to assess history-taking and communication skills among medical students. The VPs demonstrated acceptable content and face validity, suggesting that the case content, interface, and simulated environment were appropriate and realistic. However, the reliability and other forms of validity evidence, such as construct and concurrent validity, are limited. These findings may be influenced by factors such as a small sample size, variability in student preparation, limited scenario complexity, and technical challenges related to speech recognition and system responsiveness. Addressing these limitations in future research, such as by recruiting larger and more diverse cohorts, standardizing students and examiners, incorporating a greater range of progressively complex scenarios and refining the technical architecture, may enhance the robustness of the validity evidence and strengthen the overall utility of VPs as educational and assessment tools.

Despite these limitations, this work contributes to the growing body of research on immersive simulation tools in medical education. It highlights both the potential and the challenges of implementing VPs as assessment and learning tools. Further refinements in scenario design, AI integration, and the user interface are essential to maximize the pedagogical potential of VPs. With continued development, VPs hold promises as scalable and effective tools for preparing students to navigate real-world clinical encounters with competence and confidence.

Acknowledgements
The authors would like to thank all the participants and volunteers for the time and effort spent on the study.

Authors' contributions
S.K.A., N.S.R., R.H., A.A. and P.S.S. conceptualized and supervised the project. Both S.K.A. and N.A.M.S. developed the methodology and significantly prepared and edited the manuscript. N.A.M.S. collected the data and completed the statistical analysis.

Funding
This research project was funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme [FRGS/1/2019/SKK03/UPM/02/1].

Data availability
The data supporting the findings of this study are not openly available owing to confidentiality and privacy considerations but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee for Research Involving Human Subjects, Universiti Putra Malaysia which is guided in its stance and decisions by the principles expressed in the Declaration of Helsinki (reference no: JKEUPM-2020-415). The Dean of the Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, granted permission to conduct the research. All participants provided informed consent prior to participation in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 August 2025 / Accepted: 19 December 2025
Published online: 19 January 2026

References
1. Altun N, Oflaz L, Özkök G, Özzaim AÖ, Oflaz A. The role of effective communication in education management. Online J New Horizons Educ. 2024;14(1):166–73.
2. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;68(6):443–53. https://doi.org/10.1097/00001888-199306000-00002.
3. Kaplonyi J, Bowles KA, Nestel D, Kiegaldie D, Maloney S, Haines T, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ. 2017;51(12):1209–19. https://doi.org/10.1111/medu.13387.
4. Lay-Khim G, Bit-Lian Y. Experience toward simulated patient-based simulation session: an integrative literature review. Educ Med J. 2019;11(3):5–21. https://doi.org/10.21315/eimj2019.11.3.2.
5. Romancenco A, Saratila I, Ababii I, Rojnoveanu G, Dandara O, Spinei L. Bridging theory and practice: enhancing medical education through simulation-based training methods. Mold J Health Sci. 2024;11(2):68–73. https://doi.org/10.52645/mjhs.2024.2.09.
6. Mayer A, Da Silva Domingues V, Hege I, Kononowicz AA, Larrosa M, Martínez-Jarreta B, et al. Planning a collection of virtual patients to train clinical reasoning: a blueprint representative of the European population. Int J Environ Res Public Health. 2022;19(10):6175. https://doi.org/10.3390/ijerph19106175.
7. Macnamara AF, Bird K, Rigby A, Sathyapalan T, Hepburn D. High-fidelity simulation and virtual reality: an evaluation of medical students' experiences. BMJ Simul Technol Enhanc Learn. 2021;7(6):528–35. https://doi.org/10.1136/bmjstel-2020-000625.
8. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Med Educ. 2022;22(1):365. https://doi.org/10.1186/s12909-022-03410-x.
9. Jiang H, Vimalesvaran S, Wang JK, Lim KB, Mogali SR, Car LT. Virtual reality in medical students' education: scoping review. JMIR Med Educ. 2022;8(1):e34860. https://doi.org/10.2196/34860.
10. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res. 2019;21(7):e14676. https://doi.org/10.2196/14676.
11. Rothlind E, Fors U, Salminen H, Wändell P, Ekblad S. Virtual patients reflecting the clinical reality of primary care: a useful tool to improve cultural competence. BMC Med Educ. 2021;21(1):270. https://doi.org/10.1186/s12909-021-02701-z.
12. Weideman YL, Young L, Lockhart JS, Grund FJ, Fridline MM, Panas M. Strengthening cultural competence in prenatal care with a virtual community: Building capacity through collaboration. J Prof Nurs. 2016;32(5S):S48–53. https://doi.org/10.1016/j.profnurs.2016.03.004.
13. Mohd Salim NA, Roslan NS, Hod R, Zakaria SF, Adam SK. Exploring critical components of physician–patient communication: a qualitative study of Lay and professional perspectives. Healthcare. 2023;11(2):162. https://doi.org/10.3390/healthcare11020162.

14. Bansal G, Rajgopal K, Chamola V, Xiong Z, Niyato D. Healthcare in metaverse: a survey on current metaverse applications in healthcare. IEEE Access. 2022;10:119914–46. https://doi.org/10.1109/ACCESS.2022.3219845.
15. Nsaif WS, Salih HM, Saleh HH, Al-Nuaimi BT. Chatbot development: framework, platform, and assessment metrics. Eurasia Proc Sci Technol Eng Math. 2024;27:50–62.
16. Abu Shawar B, Atwell E. Chatbots: are they truly useful? J Lang Technol Comput Linguist. 2007;22(1):29–49. https://doi.org/10.21248/jlcl.22.2007.88.
17. Alazzam BA, Alkhatib M, Shaalan K. Artificial intelligence chatbots: a survey of classical versus deep machine learning techniques. Inf Sci Lett. 2023;12(4):1217–33. https://doi.org/10.18576/isl/120437.
18. Collins C, Dennehy D, Conboy K, Mikalef P. Artificial intelligence in information systems research: a systematic literature review and research agenda. Int J Inf Manag. 2021;60:102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383.
19. Zamanzadeh V, Ghahramanian A, Rassouli M, Abbaszadeh A, Alavi-Majd H, Nikanfar AR. Design and implementation content validity study: development of an instrument for measuring patient-centered communication. J Caring Sci. 2015;4(2):165–78. https://doi.org/10.15171/jcs.2015.017.
20. Alsalamah A, Campo R, Tanos V, Grimbizis G, Van Belle Y, Hood K, et al. Face and content validity of the virtual reality simulator 'ScanTrainer®'. Gynecol Surg. 2017;14(1):18. https://doi.org/10.1186/s10397-017-1020-6.
21. Alvarez-Lopez F, Maina MF, Saigí-Rubió F. Use of a low-cost portable 3D virtual reality gesture-mediated simulator for training and learning basic psychomotor skills in minimally invasive surgery: development and content validity study. J Med Internet Res. 2020;22(7):e17491. https://doi.org/10.2196/17491.
22. Alzahrani T, Haddad R, Alkhayal A, Delisle J, Drudi L, Gotlieb W, et al. Validation of the Da Vinci surgical skill simulator across three surgical disciplines: a pilot study. Can Urol Assoc J. 2013;7(7–8):E520–9. https://doi.org/10.5489/cuaj.419.
23. Klemenc-Ketis Z, Kersnik J. New virtual case-based assessment method for decision making in undergraduate students: a scale development and validation. BMC Med Educ. 2013;13:160. https://doi.org/10.1186/1472-6920-13-160.
24. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health. 2007;30(4):459–67. https://doi.org/10.1002/nur.20199.
25. Yusoff MSB. ABC of content validation and content validity index calculation. Educ Med J. 2019;11(2):49–54. https://doi.org/10.21315/eimj2019.11.2.6.
26. Hung AJ, Zehnder P, Patil MB, Cai J, Ng CK, Aron M, et al. Face, content and construct validity of a novel robotic surgery simulator. J Urol. 2011;186(3):1019–24. https://doi.org/10.1016/j.juro.2011.04.064.
27. Moores KL, Jones GL, Radley SC. Development of an instrument to measure face validity, feasibility and utility of patient questionnaire use during health care: the QQ-10. Int J Qual Health Care. 2012;24(5):517–24. https://doi.org/10.1093/intqhc/mzs051.
28. Yusoff MSB. ABC of response process validation and face validity index calculation. Educ Med J. 2019;11(3):55–61. https://doi.org/10.21315/eimj2019.11.3.6.
29. Sullivan GM. A primer on the validity of assessment instruments. J Grad Med Educ. 2011;3(2):119–20. https://doi.org/10.4300/JGME-D-11-00075.1.
30. Colaco HB, Hughes K, Pearse E, Arnander M, Tennent D. Construct validity, assessment of the learning curve, and experience of using a low-cost arthroscopic surgical simulator. J Surg Educ. 2017;74(1):47–54. https://doi.org/10.1016/j.jsurg.2016.07.006.
31. Cohen RJ, Swerdlik ME. Psychological testing and assessment: an introduction to tests and measurement. 9th ed. New York: McGraw-Hill Education; 2018.
32. Ekolu SO, Quainoo H. Reliability of assessments in engineering education using cronbach's alpha, KR and split-half methods. Glob J Eng Educ. 2019;21(1):24–9.
33. Yudkowsky R, Park YS, Downing SM, editors. Assessment in health professions education. New York: Routledge; 2019.
34. Maicher K, Danforth D, Price A, Zimmerman L, Wilcox B, Liston B, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Simul Healthc. 2017;12(2):124–31. https://doi.org/10.1097/SIH.0000000000000195.
35. Sezer B, Sezer TA, Teker GT, Elcin M. Developing a virtual patient: design, usability, and learning effect in communication skills training. BMC Med Educ. 2023;23(1):891. https://doi.org/10.1186/s12909-023-04860-7.
36. Maicher KR, Stiff A, Scholl M, White M, Fosler-Lussier E, Schuler W, et al. Artificial intelligence in virtual standardized patients: combining natural Language Understanding and rule-based dialog management to improve conversational fidelity. Med Teach. 2022;45(3):279–85. https://doi.org/10.1080/0142159X.2022.2130216.
37. Mergen M, Meyerheim M, Graf N. Reviewing the current state of virtual reality integration in medical education: a scoping review protocol. Syst Rev. 2023;12(1):97. https://doi.org/10.1186/s13643-023-02266-6.
38. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J. 2019;6(3):181–5. https://doi.org/10.7861/fhj.2019-0036.
39. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health. 2008;31(2):180–91. https://doi.org/10.1002/nur.20247.
40. Botezatu M, Hult H, Tessma MK, Fors U. Virtual patient simulation: knowledge gain or knowledge loss? Med Teach. 2010;32(7):562–8. https://doi.org/10.3109/01421590903514630.
41. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011;306(9):978–88. https://doi.org/10.1001/jama.2011.1234.
42. Fink MC, Heitzmann N, Reitmeier V, Siebeck M, Fischer F, Fischer MR. Diagnosing virtual patients: the interplay between knowledge and diagnostic activities. Adv Health Sci Educ. 2023;28(4):1245–64. https://doi.org/10.1007/s10459-023-10211-4.
43. De Mattei L, Morato MQ, Sidhu V, Gautam N, Mendonca CT, Tsai A, et al. Are artificial intelligence virtual simulated patients (AI-VSP) a valid teaching modality for health professional students? Clin Simul Nurs. 2024;92:101536. https://doi.org/10.1016/j.ecns.2024.101536.
44. Guetterman TC, Sakakibara R, Baireddy S, Kron FW, Scerbo MW, Cleary JF, et al. Medical students' experiences and outcomes using a virtual human simulation to improve communication skills: mixed methods study. J Med Internet Res. 2019;21(11):e15459. https://doi.org/10.2196/15459.
45. Quail M, Brundage SB, Spitalnick J, Allen PJ, Beilby J. Student self-reported communication skills, knowledge and confidence across standardized patient, virtual and traditional clinical learning environments. BMC Med Educ. 2016;16(1):73. https://doi.org/10.1186/s12909-016-0577-5.
46. Maicher KR, Zimmerman L, Wilcox B, Liston B, Cronau H, Macerollo A, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach. 2019;41(9):1053–9. https://doi.org/10.1080/0142159X.2019.1616683.
47. Bauler TJ, Sheakley ML, Turula H, Miller L, Busha ME. A virtual patient encounter to promote integration in a preclinical case-based learning session. Med Sci Educ. 2023;33(2):331–2. https://doi.org/10.1007/s40670-023-01744-9.
48. Nuzzo A, Tran-Dinh A, Courbebaisse M, Peyre H, Plaisance P, Matet A, et al. Improved clinical communication OSCE scores after simulation-based training: results of a comparative study. PLoS ONE. 2020;15(9):e0238542. https://doi.org/10.1371/journal.pone.0238542.

Publisher's Note