



RESEARCH

Open Access



# Analysis of virtual standardized patients for assessing clinical fundamental skills of medical students: a prospective study

Xinyu Zhang<sup>1,2†</sup>, Duo Zeng<sup>2†</sup>, Xiandi Wang<sup>2</sup>, Yaoyu Fu<sup>3</sup>, Ying Han<sup>2</sup>, Manqing He<sup>2</sup>, Xiaoling Chen<sup>2</sup> and Dan Pu<sup>1,2\*</sup>

## Abstract

**Background** History-taking is an essential clinical competency for qualified doctors. The limitations of the standardized patient (SP) in taking history can be addressed by the virtual standardized patient (VSP). This paper investigates the accuracy of virtual standardized patient simulators and evaluates the applicability of the improved system's accuracy for diagnostic teaching support and performance assessment.

**Methods** Data from the application of VSP to medical residents and students were gathered for this prospective study. In a human-machine collaboration mode, students completed exams involving taking SP histories while VSP provided real-time scoring. Every participant had VSP and SP scores. Lastly, using the voice and text records as a guide, the technicians will adjust the system's intention recognition accuracy and speech recognition accuracy.

**Results** The research revealed significant differences in scoring across several iterations of VSP and SP ( $p < 0.001$ ). Across various clinical cases, there were differences in application accuracy for different versions of VSP ( $p < 0.001$ ). Among training groups, the diarrhea case showed significant differences in speech recognition accuracy ( $Z = -2.719$ ,  $p = 0.007$ ) and intent recognition accuracy ( $Z = -2.406$ ,  $p = 0.016$ ). Scoring and intent recognition accuracy improved significantly after system upgrades.

**Conclusion** VSP has a comprehensive and detailed scoring system and demonstrates good scoring accuracy, which can be a valuable tool for history-taking training.

**Keywords** Virtual standardized patient, Standardized patient, Simulation-based education, Clinical skills, History-taking

## Background

History-taking is an essential skill for becoming a competent doctor, and it is a fundamental component of work in various medical fields [1]. History-taking typically includes general data, chief complaints, present history, past medical history, family history, social history, and review of systems, etc. Since it directs subsequent exams, diagnosis, and treatment choices, gathering patient histories is the first and most important stage in identifying medical conditions [2]. Therefore, it is vital to provide

<sup>†</sup>Xinyu Zhang and Duo Zeng contributed equally to this work.

\*Correspondence:

Dan Pu  
pudan8012@wchscu.cn

<sup>1</sup>The Chinese Cochrane Center, West China Hospital, Sichuan University, Chengdu, Sichuan Province 610041, People's Republic of China

<sup>2</sup>West China Medical Simulation Center, West China Hospital, Sichuan University, Chengdu, Sichuan Province 610041, People's Republic of China

<sup>3</sup>West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan Province 610041, People's Republic of China



medical professionals with training in history-taking before they engage in clinical practice [3–6].

Currently, the most common method for teaching history-taking skills combines theoretical instruction with simulation-based education, with Standardized Patients (SP) as the primary method of simulation. SP are individuals who have undergone standardized and systematic training to accurately, realistically, and consistently portray the characteristics, psychosocial features, and emotional responses required for specific medical cases [7, 8]. Doctor-patient communication encompasses both verbal and non-verbal components, which are equally important [2, 9]. During the diagnostic process, doctors collect information from patients by observing their facial expressions and body language. Similarly, doctors use body language and facial expressions to encourage and ensure patient comfort [10, 11]. The United States first used SP for clinical teaching in the 1960s, and China adopted the practice in the 1990s [12]. SP teaching is a valuable bridge between theoretical instruction and clinical practice. It not only facilitates the simulation of authentic medical scenarios without ethical concerns but also boosts student engagement, enhances clinical communication skills, supports the acquisition of medical knowledge, and promotes a deeper grasp of abstract concepts [13].

However, the use of SP in medical education has its own challenges. The training process for SP is rigorous, time-consuming, and resource-intensive. Consequently, the availability of qualified SP is limited [14, 15]. In the process of SP teaching evaluation, the influence of subjective factors cannot be avoided [16]. The lengthy and strict training process, resulting in the scarcity of SP, makes it challenging to implement one-on-one history-taking training effectively. To address these limitations of SP, virtual standardized patient (VSP) offers a potential solution. As early as the early twenty-first century, research suggested using computers to aid in history-taking exercises [17–20], but VSP has not become widely adopted. Implementing VSP in history-taking instruction can effectively address the limitations found in SP. It reduces the lengthy training time and costs associated with training SP, allows for repetitive training [21, 22], and facilitates the assessment of teaching effectiveness [23], thereby boosting student confidence [24, 25]. It reduces the potential subjectivity of both instructors and SP, enabling a more objective and standardized evaluation [23].

We developed a VSP according to the needs, using speech recognition technology, intention recognition technology, and automatic scoring. VSP initially used sentence similarity matching and then improved to intention recognition. The article aims to explore the accuracy

of VSP and assess whether the upgraded system's accuracy can be applied to diagnostic teaching assistance and performance evaluation. This research highlights certain limitations in SP physician training and examines the application accuracy of our independently developed VSP. The goal is to establish a foundation for more effective teaching strategies.

## Methods

### VSP

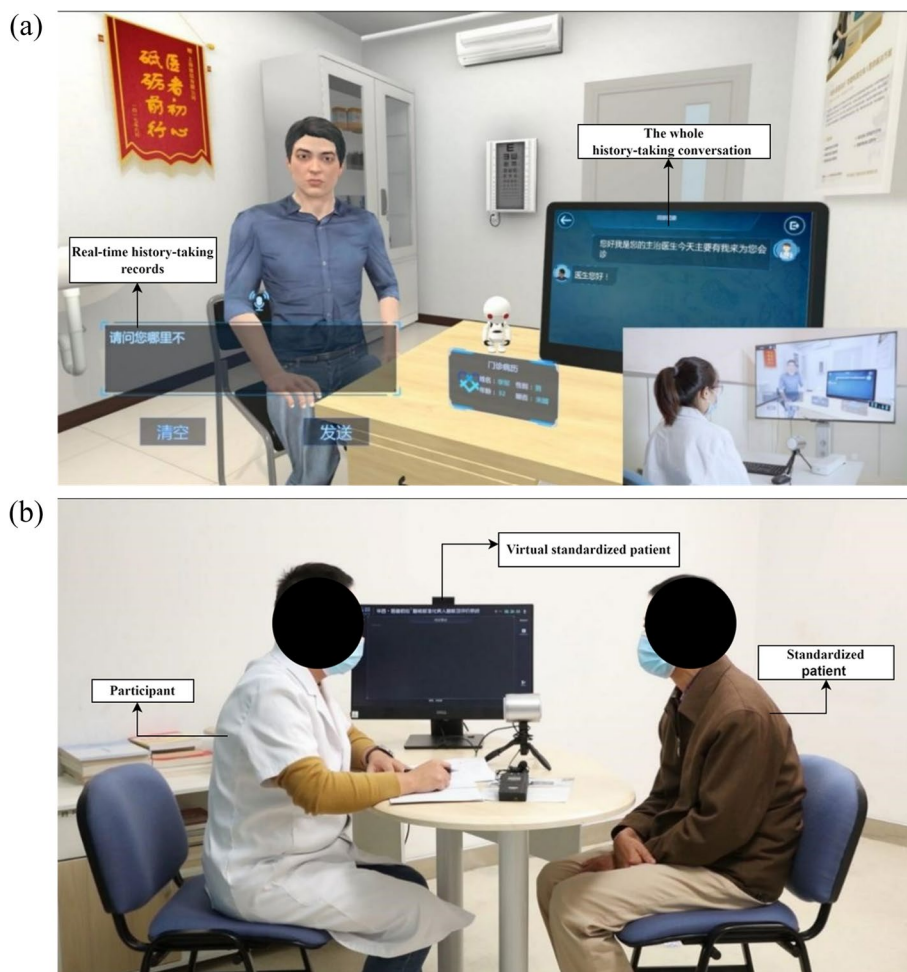
This study utilizes a virtual standardized patient history-taking system jointly developed by our institution and Shanghai Chuxin Medical Technology Co., Ltd., based on speech recognition and intent recognition technology. The system operates in both a human–computer dialogue mode and a human–computer collaborative mode, as detailed in Fig. 1. This research employs the latter.

The system first converts the spoken dialogue into text. Subsequently, the sentences are dissected, breaking them down into phrases. Following part-of-speech recognition and classification, these sentences are compared to the intent templates kept in the intent library to provide assessments and comments. After gathering all the data, the system performs self-learning to adjust the corpus [26]. The specific process is illustrated in Fig. 2.

VSP underwent a general system replacement during the experiment, with VSP 1.0 being the old version and VSP 2.0 and VSP 3.0 being the new version. VSP 1.0 compares sentences and keywords with standard statements, for the old version of the system. It is considered the same statement when the sentences are similar to the standard statements [27]. For the updated version of the system, VSP 2.0 divides the sentences into phrases, categorizes the phrases, and matches the intent templates of the phrases [28]. VSP 3.0 is the version of VSP 2.0 self-learning optimization [29–32]. The differences between the old and new versions are shown in Fig. 3.

### Design, setting and subjects

We adopted a prospective study design. The Biomedical Ethics Committee of West China Hospital, Sichuan University approved this study (Approval 2019 No.1071). In this study, we applied different versions of VSP to assess the clinical performance of medical students with no prior clinical experience and residents, using a human–computer collaborative model for evaluation. In the study population, clinical medical students were recruited from the annual diagnostics course, while residents were selected from the enhanced training sessions, which covered three years. Participants willing to use VSP are recruited from these two courses, and the scoring results of history-taking can be compared with those of SP. Informed consent was obtained from all participants



**Fig. 1** Human-computer dialogue mode and human-computer collaborative mode of VSP **a** human-computer dialogue mode; **b** human-computer collaborative mode

before the tests, and they were informed that the results of this study would not affect their final course grades. All participants had previously received theoretical instruction in medical history taking.

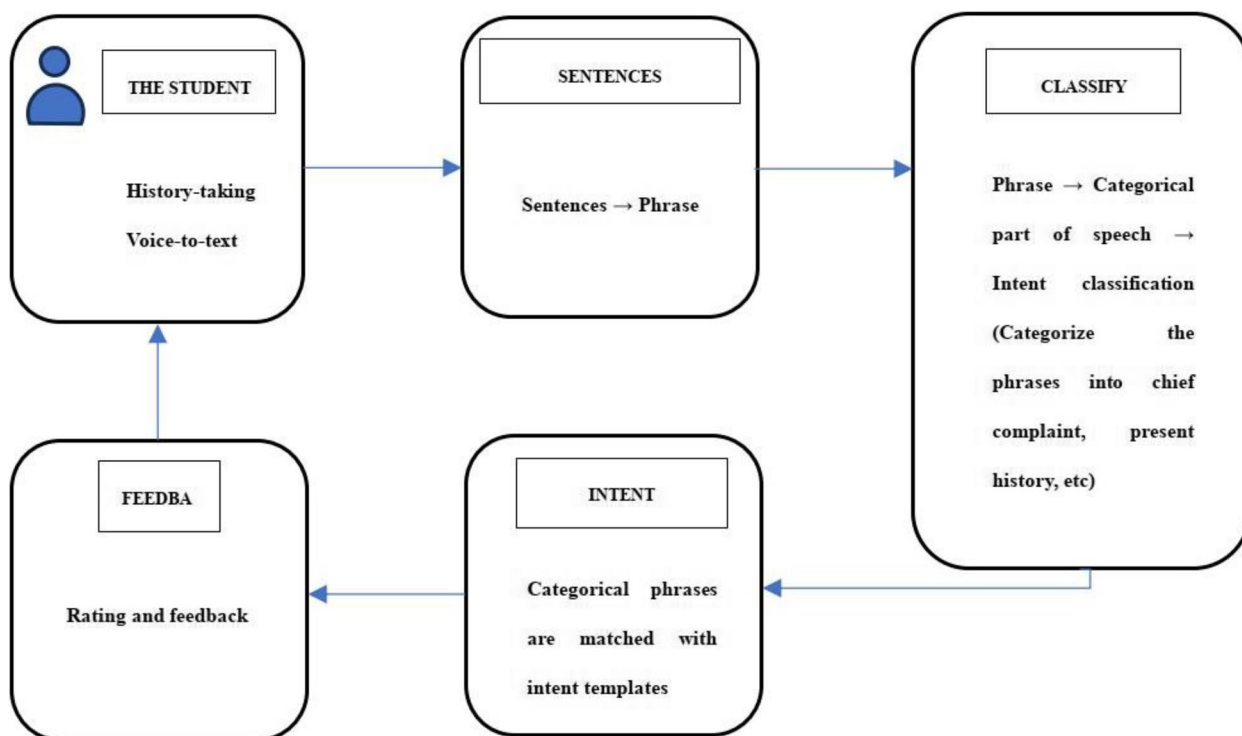
**Measurements**

In this study, the accuracy of the VSP application was determined. The application accuracy included speech recognition accuracy, intention recognition accuracy, and scoring accuracy. Speech recognition accuracy is the ratio of correctly recognized characters to the total number of characters. Intent recognition accuracy is the ratio of correctly matched phrases to the total number of phrases. The system automatically determines the accuracy of speech recognition and intent recognition, and it separates intent matches with a probability of less than 80%. The results were reviewed manually by two technicians. If their results differed, a third technician made the final decision. The score consisted of the content of

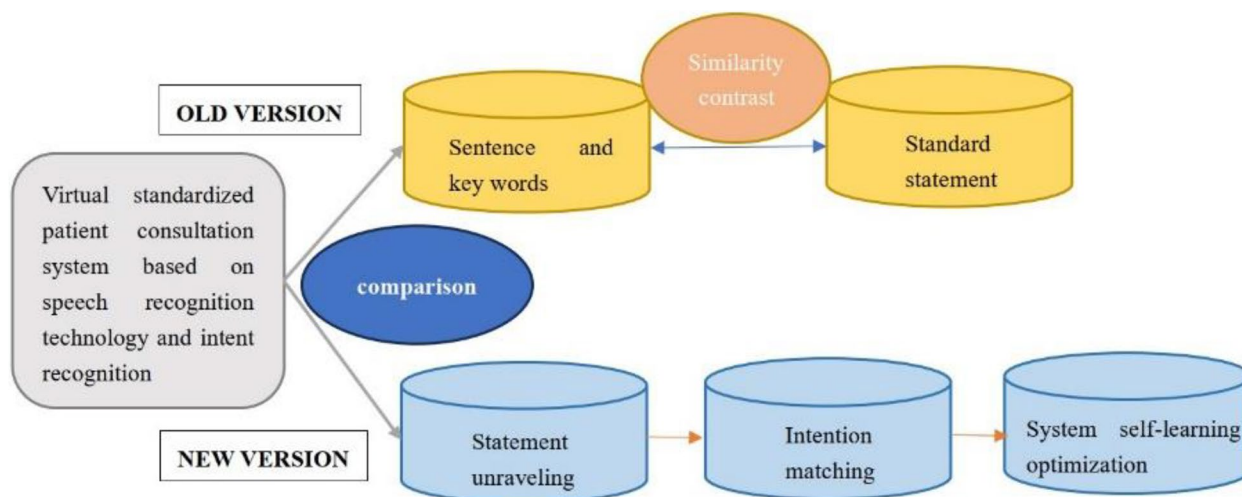
history-taking and the skill of history-taking, with a total of 70 scoring points, and the total score of each scoring point was different. The Department of Diagnostics’ multidisciplinary team established the scoring scale, which has been validated and applied for many years. The scale underwent minor modifications based on actual conditions to ensure quality control. Because SP is highly trained and experienced, we calculated the VSP’s scoring accuracy using its scores as the gold standard. Scoring accuracy was calculated as the ratio of the number of scoring points with the same VSP and SP scores to the total score points.

**Procedure**

In this study, the participants randomly selected one case from the four cases (diarrhea, syncope, palpitation, cough) during the assessment process. Throughout the assessment, SP acted as a patient and interacted with participants performing the role of doctors. The



**Fig. 2** The specific process of the virtual standardized patient history-taking system based on speech recognition technology and artificial intelligence



**Fig. 3** Comparison of old and new VSP

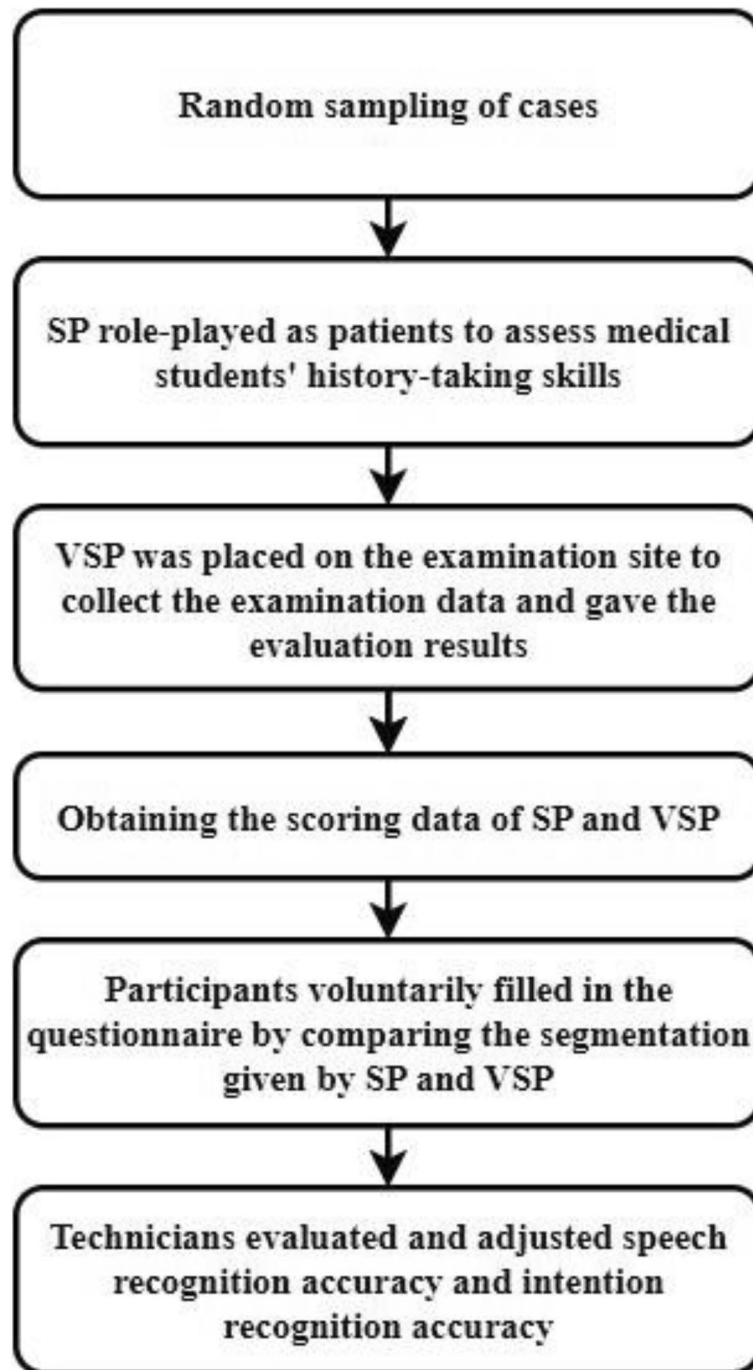
VSP was placed next to it without responding, collecting information for real-time scoring. As a result, two sets of scores were obtained from SP and VSP. After the examination, participants were invited to complete relevant questionnaires voluntarily (results are shown in the appendices). Following history-taking training with SP, both SP and VSP scores are simultaneously

obtained, and participants provide feedback about the VSP after comparing these scores. Speech and intent recognition employ mature commercial technologies, with accuracy automatically generated by the system upon the corpus. Lastly, technicians reviewed the texts and recordings to assess and adjust the accuracy provided by the system. See Fig. 4 for details.

**Data analysis**

Since the data did not meet the normal distribution, we used the Wilcoxon rank sum test when comparing the SP and VSP scores. Analysis of variance (ANOVA) was used to test the significance of the differences in the accuracy

of scoring, speech recognition, and intent recognition of VSP across different VSP versions in various cases. Kruskal–Wallis one-way ANOVA tests were used for pairwise comparisons, with post hoc analysis using the Bonferroni correction. The independent *t*-test was used to compare the accuracy of scoring, speech recognition,



**Fig. 4** The procedure of the study

and intent recognition of VSP 3.0 in different cases between medical students and residents.  $p < 0.05$  was considered statistical significance.

**Results**

**Demographic characteristics**

A total of 502 students participated in the study over the three years. Of these, 476 students were included in the final analysis, as 26 students' data were not recorded due to VSP 1.0 system problems. Among the included participants, 89 medical students used VSP 1.0, 129 used VSP 2.0, and 104 used VSP 3.0. The 154 residents used VSP 3.0. Statistics were based on different versions and randomly selected cases, as shown in Table 1:

The medical students who used different versions of VSP are at similar stages of learning history-taking, with comparable ages. Residents have more clinical experience than medical students in the same stage of training.

**Comparison of history-taking scores given by SP and VSP**

The t-test revealed significant differences in the scores given by SP and VSP for both medical students ( $Z = -8.194, p < 0.05; Z = -9.864, p < 0.05; Z = -8.867, p < 0.05$ ) and residents ( $Z = -10.773, p < 0.05$ ). Generally, VSP scores were lower than SP scores. The score distribution was skewed, mainly in the high-score range, as shown in Table 2.

**Table 1** Distribution of subjects ( $n = 476$ )

Cases	VSP 1.0	VSP 2.0	VSP 3.0	
	Medical students	Medical students	Medical students	Residents
Diarrhea	19	30	27	39
Syncope	16	27	26	37
Palpitations	35	48	26	39
Cough	19	24	25	39
Total	89	129	104	154

**Table 2** Comparison of SP and VSP history-taking Scores ( $n = 476$ )

Version		N	Mean(SD)	Median(Min-Max)	Z <sup>a</sup>	p
VSP 1.0	SP	89	128.52(9.80)	131(98-141)	-8.194	0.001
	VSP	89	88.85(7.51)	90.5(44.5-101.5)		
VSP 2.0	SP	129	136.29(7.85)	138(105-148)	-9.864	0.001
	VSP	129	119.05(7.53)	120(91-134)		
VSP 3.0 (Medical students)	SP	104	137.43(8.42)	140(105-150)	-8.867	0.001
	VSP	104	121.05(7.91)	123(92-136)		
VSP 3.0 (Residents)	SP	154	126.87(12.40)	129(96-150)	-10.773	0.001
	VSP	154	111.08(10.76)	112(85-130)		

<sup>a</sup> Due to the data not following a normal distribution, the Wilcoxon test was chosen

**Comparison of VSP application accuracy**

Our study examined four distinct medical scenarios, comparing the application accuracy of various VSP versions and determining whether there are variations in accuracy while using VSP 3.0 with medical students and residents.

**Comparison of VSP application accuracy in diarrhea cases**

The results indicated significant differences in the application accuracy ( $H = 42.424, p < 0.001; H = 27.220, p < 0.001; H = 44.135, p < 0.001$ ) among the three versions of the VSP system. Multiple mean comparisons revealed significant differences in scoring accuracy between VSP 1.0 and VSP 2.0 ( $p < 0.001$ ), VSP 1.0 and VSP 3.0 ( $p < 0.001$ ). The speech recognition accuracy between VSP 1.0 and VSP 3.0 ( $p < 0.001$ ), and VSP 2.0 and VSP 3.0 ( $p < 0.001$ ) was significantly different. Intent recognition accuracy was significantly different between VSP 1.0 and VSP 2.0 ( $p < 0.001$ ), VSP 1.0 and VSP 3.0 ( $p < 0.001$ ). The results are presented in Table 3.

When instructing medical students and residents in history-taking using VSP 3.0, a t-test was employed to determine whether there were any significant differences in application accuracy between the two groups. The results showed significant differences in speech recognition accuracy ( $Z = -2.719, p = 0.007$ ) and intent recognition accuracy ( $Z = -2.406, p = 0.016$ ). The results are presented in Table 4.

**Comparison of VSP application accuracy in syncope cases**

There were significant differences in the application accuracy ( $H = 34.506, p < 0.001; H = 27.233, p < 0.001; H = 38.485, p < 0.001$ ). Multiple mean comparison results showed significant differences in scoring accuracy between VSP 1.0 and VSP 2.0 ( $p < 0.001$ ), as well as between VSP 1.0 and VSP 3.0 ( $p < 0.001$ ). Significant differences were observed in speech recognition accuracy between VSP 1.0 and VSP 2.0 ( $p < 0.001$ ),

**Table 3** Comparison of VSP application accuracy in diarrhea cases (n = 76)

VSP version	Scoring accuracy (%)		Speech recognition accuracy (%)			Intent recognition accuracy (%)		
	Median(P25,P75)	Rank Sum test	Median(P25,P75)	Rank Sum test	Median(P25,P75)	Rank Sum test	H-Value <sup>g</sup>	p-Value
		H-Value <sup>g</sup>		H-Value <sup>g</sup>		H-Value <sup>g</sup>		
VSP 1.0	67.88(65.39,71.72) <sup>a</sup>	42.424	95.11(94.03,96.77) <sup>c</sup>	27.220	77.31(76.41,78.11) <sup>e</sup>	44.135		0.001
VSP 2.0	87.85(85.59,89.48) <sup>b</sup>		95.05(94.33,95.74) <sup>c</sup>		82.59(81.16,83.30) <sup>f</sup>			
VSP 3.0	88.11(86.62,89.44) <sup>b</sup>		96.79(96.34,97.33) <sup>d</sup>		83.37(82.33,83.79) <sup>f</sup>			

<sup>a, b, c, d, e, f</sup> If two groups share the same letter, it indicates no significant difference; if there are no common letters, it signifies a significant difference

<sup>g</sup> Due to the data not following a normal distribution or not satisfying homogeneity of variance, Kruskal–Wallis H analysis was chosen

between VSP 1.0 and VSP 3.0 ( $p=0.016$ ), and between VSP 2.0 and VSP 3.0 ( $p=0.019$ ). Intent recognition accuracy exhibited significant differences between VSP 1.0 and VSP 2.0 ( $p<0.001$ ), between VSP 1.0 and VSP 3.0 ( $p<0.001$ ), and between VSP 2.0 and VSP 3.0 ( $p=0.036$ ). The results are presented in Table 5.

There were no significant differences in application accuracy ( $Z=-0.426, p=0.670$ ;  $Z=-0.216, p=0.829$ ;  $Z=-0.035, p=0.972$ ) between medical students and residents using VSP 3.0 in syncope cases. The results are presented in Table 6.

**Comparison of VSP application accuracy in palpitation cases**

There were significant differences in the application accuracy ( $H=71.858, p<0.001$ ;  $H=23.986, p<0.001$ ;  $H=77.121, p<0.001$ ). Multiple mean comparison results showed significant differences in scoring accuracy between VSP 1.0 and VSP 2.0 ( $p<0.001$ ), as well as between VSP 1.0 and VSP 3.0 ( $p<0.001$ ). Significant differences were observed in speech recognition accuracy between VSP 1.0 and VSP 2.0 ( $p=0.011$ ), VSP 1.0 and VSP 3.0 ( $p<0.001$ ), and between VSP 2.0 and VSP 3.0 ( $p=0.035$ ). Intent recognition accuracy exhibited

significant differences between VSP 1.0 and VSP 2.0 ( $p<0.001$ ), VSP 1.0 and VSP 3.0 ( $p<0.001$ ), and between VSP 2.0 and VSP 3.0 ( $p=0.033$ ). The results are presented in Table 7.

The results showed no significant differences in application accuracy ( $t=1.055, p=0.132$ ;  $t=0.138, p=0.068$ ;  $t=-0.872, p=0.557$ ) when using VSP 3.0 for teaching medical students and residents in palpitation cases. The results are presented in Table 8.

**Comparison of VSP application accuracy in cough cases**

There were significant differences in the application accuracy ( $H=40.521, p<0.001$ ;  $H=18.961, p<0.001$ ;  $F=235.851, p<0.001$ ). Multiple mean comparison results indicated significant differences in scoring accuracy between VSP 1.0 and VSP 2.0 ( $p<0.001$ ), as well as between VSP 1.0 and VSP 3.0 ( $p<0.001$ ). Significant differences were observed in speech recognition accuracy between VSP 1.0 and VSP 2.0 ( $p<0.001$ ), as well as between VSP 1.0 and VSP 3.0 ( $p=0.011$ ). Intent recognition accuracy exhibited significant differences between VSP 1.0 and VSP 2.0 ( $p<0.001$ ), between VSP

**Table 4** Comparison of the accuracy of VSP application in different groups of diarrhea cases ( $n=66$ )

Training population	Scoring accuracy (%)			Speech recognition accuracy (%)			Intent recognition accuracy (%)		
	mean (SD)	t-Test		Median(P25,P75)	Rank Sum test		Median(P25,P75)	Rank Sum test	
		Z-Value	p-Value		Z-Value	p-Value		Z-Value	p-Value
Medical students	87.99(2.00)	1.220	0.353	96.79(96.34,97.33)	-2.719	0.007	83.37(82.33,83.79)	-2.406	0.016
Residents	87.30(2.45)			96.26(95.64,96.98)			83.96(82.60,85.04)		

**Table 5** Comparison of VSP application accuracy in syncope cases ( $n=69$ )

VSP version	Scoring accuracy (%)			Speech recognition accuracy (%)			Intent recognition accuracy (%)		
	Median(P25, P75)	Rank Sum test		Median(P25,P75)	Rank Sum test		Median(P25,P75)	Rank Sum test	
		H-Value	p-Value		H-Value	p-Value		H-Value	p-Value
VSP 1.0	79.04(75.31,81.45) <sup>a</sup>	34.506	0.001	97.67(97.28,97.78) <sup>c</sup>	27.233	0.001	79.19(78.91,79.50) <sup>f</sup>	38.485	0.001
VSP 2.0	87.59(84.85,88.98) <sup>b</sup>			95.63(94.70,96.16) <sup>d</sup>			82.04(80.02,84.41) <sup>g</sup>		
VSP 3.0	88.63(87.07,89.61) <sup>b</sup>			96.26(95.57,97.17) <sup>e</sup>			83.49(82.91,85.09) <sup>h</sup>		

<sup>a, b, c, d, e, f, g, h</sup> If two groups share the same letter, it indicates no significant difference; if there are no common letters, it signifies a significant difference

**Table 6** Comparison of the accuracy of VSP application in different groups of syncope cases ( $n=63$ )

Training population	Scoring accuracy (%)			Speech recognition accuracy (%)			Intent recognition accuracy (%)		
	Median(P25, P75)	Rank Sum test		Median(P25,P75)	Rank Sum test		Median(P25,P75)	Rank Sum test	
		Z-Value	p-Value		Z-Value	p-Value		Z-Value	p-Value
Medical students	88.63(87.07,89.61)	-0.426	0.670	96.26(95.57,97.17)	-0.216	0.829	83.49(82.91,85.09)	-0.035	0.972
Residents	88.24(86.67,89.62)			96.50(95.55,96.92)			83.68(82.63,85.35)		



**Table 7** Comparison of VSP application accuracy in palpitation cases ( $n = 109$ )

VSP version	Scoring accuracy (%)			Speech recognition accuracy (%)			Intent recognition accuracy (%)		
	Median(P25,P75)	Rank Sum test		Median(P25,P75)	Rank Sum test		Median(P25,P75)	Rank Sum test	
		H- Value	p-Value		H- Value	p- Value		H- Value	p- Value
VSP 1.0	68.46(66.41,71.54) <sup>a</sup>	71.858	0.001	94.41(93.39,95.97) <sup>c</sup>	23.986	0.001	75.02(74.19,77.01) <sup>f</sup>	77.121	0.001
VSP 2.0	87.14(85.56,88.97) <sup>b</sup>			95.70(94.87,96.32) <sup>d</sup>			81.70(80.58,83.75) <sup>g</sup>		
VSP 3.0	88.24(87.12,89.06) <sup>b</sup>			96.27(95.61,97.48) <sup>e</sup>			83.44(82.88,84.93) <sup>h</sup>		

a, b, c, d, e, f, g, h If two groups share the same letter, it indicates no significant difference; if there are no common letters, it signifies a significant difference

**Table 8** Comparison of the accuracy of VSP application in different groups of palpitation cases ( $n = 65$ )

Training population	Scoring accuracy (%)			Speech recognition accuracy (%)			Intent recognition accuracy (%)		
	mean (SD)	t-Test		mean (SD)	t-Test		mean (SD)	t-Test	
		t- Value	p-Value		t- Value	p- Value		t- Value	p- Value
Medical students	87.98(1.95)	1.055	0.132	96.38(1.05)	0.138	0.068	83.79(1.18)	-0.872	0.557
Residents	87.34(2.67)			96.35(0.79)			84.04(1.10)		

1.0 and VSP 3.0 ( $p < 0.001$ ), and between VSP 2.0 and VSP 3.0 ( $p < 0.001$ ). The results are presented in Table 9.

There were no significant differences in application accuracy ( $t = 0.276$ ,  $p = 0.241$ ;  $t = -4.933$ ,  $p = 0.186$ ;  $t = -0.486$ ,  $p = 0.309$ ) when using VSP 3.0 for teaching medical students and residents in cough cases. The results are presented in Table 10.

#### Changes in the accuracy of different cases

We conducted an analysis and comparison of scoring accuracy, speech recognition accuracy, and intent recognition accuracy (Fig. 5). Results showed that both scoring accuracy and intent recognition accuracy increased with the upgrade of the VSP version, and the standard deviation decreased. However, the trend in speech recognition accuracy varied depending on the cases. In the VSP 1.0 version, the syncope cases showed the best accuracy in speech recognition and intent recognition, followed by diarrhea, palpitations, and coughing. In the VSP 2.0 and VSP 3.0 versions, scoring and intent recognition accuracy were nearly identical for all four cases.

#### Discussion

We explore the accuracy of our self-developed VSP simulator for assessing history-taking skills. While intent recognition and score accuracy have increased after updates and optimizations, speech recognition accuracy has continuously maintained a high level. The VSP application accuracy has stabilized after optimization and updating, continuously reaching high levels in various scenarios.

The application accuracy of VSP does not vary depending on the population.

It is clear from the statistics in Table 2. that VSP scores are generally lower than SP scores. This finding aligns with the results of a study by Fink and others [33]. Fink et al. attributed this to the lower subjects' interest, reduced appraisal of motivational value, and decreased quantity of evidence generation reported for VPs. However, our VSP did not engage in human-computer dialogue, so we believe the reason is different. Based on the analysis of this study, the reasons may be attributed to the overall operational processes and assessment methods of the VSP system. The accuracy of the system's voice recognition and intention recognition might have affected the scoring accuracy since VSP will first translate the speech into text, then perform intention recognition, and finally provide the score. The lower score of VSP compared to SP may result from the recognition and classification error of speech and intention. Therefore, our study further explored the scoring accuracy, speech recognition accuracy, and intent recognition accuracy of VSP.

Considering the potential confounding effects of different case content, we conducted separate analyses of scoring accuracy, speech recognition accuracy, and intent recognition accuracy for each of the four cases: diarrhea, syncope, palpitations, and cough. All versions of VSP used these same four cases. We analyzed the scoring accuracy, speech recognition accuracy, and intent recognition accuracy of different VSP versions in these cases. The results all showed significant differences.

We also looked at any discrepancies in accuracy between medical students and residents using VSP 3.0.

**Table 9** Comparison of VSP application accuracy in cough cases (n = 68)

VSP version	Scoring accuracy (%)		Speech recognition accuracy (%)		Intent recognition accuracy (%)	
	Median(P25, P75)	Rank Sum test H-Value p-Value	Median(P25, P75)	Rank Sum test H-Value p-Value	mean (SD)	Analysis of Variance F-Value p-Value
VSP 1.0	67.88(65.60,71.74) <sup>a</sup>	40.521 0.001	94.11(93.56,94.78) <sup>c</sup>	18.961 0.001	74.09(1.94) <sup>e</sup>	235.851 0.001
VSP 2.0	88.04(85.47,89.81) <sup>b</sup>		96.00(95.05,96.57) <sup>d</sup>		82.44(1.64) <sup>f</sup>	
VSP 3.0	87.77(86.97,89.44) <sup>b</sup>		95.24(94.67,96.09) <sup>d</sup>		83.88(1.12) <sup>g</sup>	

<sup>a, b, c, d, e, f, g</sup> If two groups share the same letter, it indicates no significant difference; if there are no common letters, it signifies a significant difference

**Table 10** Comparison of the accuracy of VSP application in different groups of cough cases ( $n = 64$ )

Training population	Scoring accuracy (%)			Speech recognition accuracy (%)			Intent recognition accuracy (%)		
	mean (SD)	t-Test		mean (SD)	t-Test		mean (SD)	t-Test	
		t-Value	p-Value		t-Value	p-Value		t-Value	p-Value
Medical students	87.92(2.03)	0.276	0.241	95.25(1.28)	-4.933	0.186	83.88(1.12)	-0.486	0.309
Residents	87.76(2.44)			96.57(0.86)			84.02(1.01)		

Among the results of comparing different groups, significant differences were observed only in the case of diarrhea, where speech recognition accuracy and intent recognition accuracy showed differences. The reason may be that medical students and residents conducted history-taking in different ways. Medical students, lacking clinical experience, tend to follow a standardized model provided by professors, making their approach easily recognizable by the system. In contrast, residents have some clinical experience, which leads to various inquiry styles that pose certain identification challenges. Furthermore, there are regional and dialectal accent variances in Chinese, which contributes to some degree of mistakes in the voice recognition system.

Based on the scoring accuracy data, the latest version of VSP achieved an accuracy rate of 85.40–89.62%, which aligns with similar research findings. In a study by William and colleagues [34] response accuracy ranged from 84 to 88%, and in Maicher et al.'s study [35], response accuracy ranged from 79 to 86%. The construction of this system has been relatively successful. However, future work should focus on enriching the synonym database and improving accuracy. The results of pairwise comparisons indicate a significant improvement in scoring accuracy with the newer system versions, i.e. VSP 2.0 and VSP 3.0. However, when compared with VSP 2.0, VSP 3.0 showed no improvement in scoring accuracy, indicating that the system's self-learning functionality has a limited impact on enhancing scoring accuracy. The reason might be insufficient data in the collected corpus and insufficient time for the machine's self-learning. It remains uncertain whether the self-learning feature of VSP has any impact on scoring accuracy. Further research is needed to confirm whether the self-learning functionality of VSP affects scoring accuracy.

In the four medical cases, speech recognition accuracy is relatively high. After pairwise comparisons, no specific patterns causing significant differences in speech recognition accuracy were observed in the data. There are possible reasons for this phenomenon. Firstly, regional and ethnic differences may contribute to distinct accents, especially when the system is designed for standard Mandarin. Secondly, speaking at a fast pace could cause the

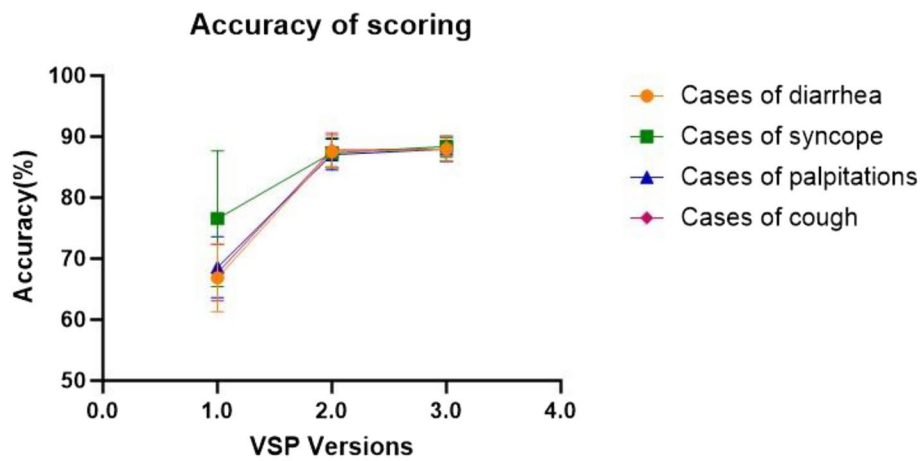
system to have difficulty accurately capturing the spoken words. Lastly, the system may fail to recognize or accurately identify sentence breaks, which could also be a contributing factor. When speech-to-text conversion fails to accurately convey the intended meaning, the system responds with errors or fails to respond. This aligns with the findings of Kammoun et al. [36], whose system automatically moves to the next section if it cannot accurately recognize the speech. In the future, adjustments can be made to the system to optimize the speech recognition section by customizing response time intervals for each individual.

Overall, intent recognition accuracy has been consistently improving. This suggests that VSP 2.0 successfully addressed the issue of intent recognition accuracy compared to VSP 1.0, and its self-learning feature provides an advantage in enhancing intent recognition accuracy. Based on the research results, it can be inferred that VSP's intent recognition accuracy does not vary with different experience groups. Future research should include a more diverse range of participants to validate these findings.

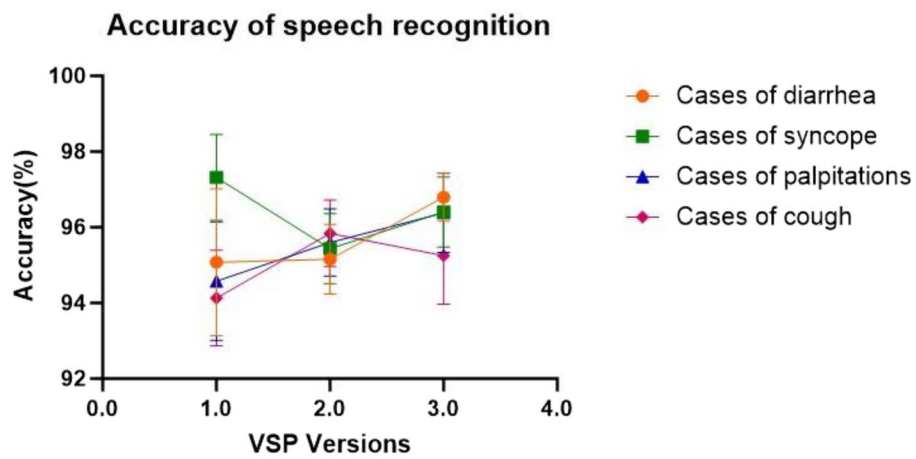
The findings (Fig. 5) indicate that scoring accuracy and intent recognition accuracy improve with the upgrade of the VSP version. However, speech recognition accuracy varies across different cases. This discrepancy can be attributed to factors mentioned earlier, such as the subject's accent, speaking speed, and sentence breakage, posing challenges for VSP recognition. In VSP 1.0, there was a notable standard deviation in application accuracy, with diarrhea cases showing the highest accuracy. This variation may be linked to VSP 1.0's slight instability and differing word recognition accuracy across cases. The scoring process involves speech recognition followed by intent recognition, leading to relatively consistent results in scoring accuracy, speech recognition accuracy, and intent recognition accuracy in VSP 1.0.

### Limitation

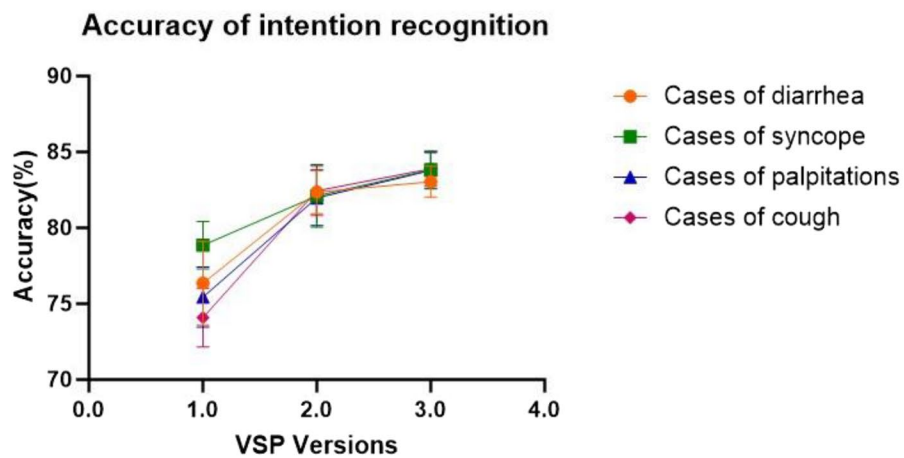
This study only utilized the system's examination mode, focusing solely on speech conversion and score feedback. The system we employed has an additional



(a)



(b)



(c)

**Fig. 5** The accuracy trend chart and the maximum accuracy was 100%. **a** is the accuracy of scoring; **b** is the accuracy of speech recognition; **c** is the accuracy of intention recognition

human–computer interaction mode that can be used for student history-taking training, which we did not explore in this study. Moreover, the system’s comprehension of voice text and response accuracy were not examined in this study. These aspects can be studied in future research. Furthermore, we only examined a few key metrics, including VSP score accuracy, speech recognition accuracy, and intent recognition accuracy, without discussing all the metrics. Additionally, this study only included medical students and residents. Extra variables for various demographics should be considered for analysis to investigate the correctness of the application of VSP in various population groupings.

## Conclusion

VSP proves to be a feasible way to train history-taking skills. This study describes the scoring process of our self-developed VSP and reveals its commendable application accuracy. The upgrading and the self-learning function of the system have played a role in improving the stability and accuracy of VSP. At this point, the accuracy of VSP 3.0 has reached the level required for the history-taking training auxiliary tool, opening up possibilities for integrating diagnostic training tools into clinical education, and effectively addressing the shortage of opportunities for students in SP training. In the future, continuous optimization of VSP will position it as a reliable training and assessment tool, fostering students’ independent learning abilities in classroom teaching.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12909-024-05982-2>.

Supplementary Material 1.

## Acknowledgements

We would like to thank all the participants and colleagues involved in this study.

## Authors’ contributions

XZ, DZ, and DP conceptualized the study. MH, YH, XC, and XW implemented the experiment. DZ and XZ collected and analyzed the data. XZ, DZ, and XW wrote the original draft. YF revised and polished the manuscript. All authors contributed to revising the manuscript and reviewing the manuscript.

## Funding

This study was supported by a Research project of the New Century Higher Education Teaching Reform Project (the ninth phase) of Sichuan University (SCU9334), the Medical Simulation Education Research Project of the National Center for Medical Education Development (2021MNYB04), and the Experimental Technology Research Program of Sichuan University (SCU2023005).

## Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

## Declarations

### Ethics approval and consent to participate

The study protocol was approved by the University of West China Hospital of Sichuan University Research Ethics Board (Approval 2019 No.1071). Each participant provided written consent before entering the study. The study was designed and conducted according to the Declaration of Helsinki and China’s competent laws and regulations.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 25 February 2024 Accepted: 3 September 2024

Published online: 10 September 2024

## References

1. Tawanwongsri W, Phenwan T. Reflective and feedback performances on Thai medical students’ patient history-taking skills. *BMC Med Educ.* 2019;19(1):141.
2. Vogel D, Meyer M, Harendza S. Verbal and non-verbal communication skills including empathy during history taking of undergraduate medical students. *BMC Med Educ.* 2018;18(1):157.
3. Alharbi L, Almoallim H. History-Taking Skills in Rheumatology. In: Almoallim H, Cheikh M, editors. *Skills in Rheumatology*. Singapore: Springer; 2021. p. 3–16.
4. Kantar A, Marchant JM, Song WJ, Shields MD, Chatziparasidis G, Zacharasiewicz A, Moeller A, Chang AB. History taking as a diagnostic tool in children with chronic cough. *Front Pediatr.* 2022;10:850912.
5. Steinkellner C, Schlömmner C, Dünser M. Medical history taking and clinical examination in emergency and intensive care medicine. *Med Klin Intensivmed Notfmed.* 2020;115(7):530–8.
6. Altshuler L, Wilhite JA, Hardowar K, Crowe R, Hanley K, Kalet A, Zabar S, Gillespie C, Ark T. Understanding medical student paths to communication skills expertise using latent profile analysis. *Med Teach.* 2023;45(10):1140–7.
7. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *AAMC. Acad Med.* 1993;68(6):443–51. Discussion 451–443.
8. Gillette C, Stanton RB, Rockich-Winston N, Rudolph M, Anderson HG Jr. Cost-effectiveness of using standardized patients to assess student-pharmacist communication skills. *Am J Pharm Educ.* 2017;81(10):73–9.
9. Bagacean C, Cousin I, Ubertini A-H, El Yacoubi El Idrissi M, Bordron A, Mercadie L, Garcia LC, Ianotto J-C, De Vries P, Berthou C. Simulated patient and role play methodologies for communication skills and empathy training of undergraduate medical students. *BMC Med Educ.* 2020;20(1):491.
10. Ishikawa H, Hashimoto H, Kinoshita M, Fujimori S, Shimizu T, Yano E. Evaluating medical students’ non-verbal communication during the objective structured clinical examination. *Med Educ.* 2006;40(12):1180–7.
11. Mehrabian A, Ferris SR. INFERENCE OF ATTITUDES FROM NONVERBAL COMMUNICATION IN 2 CHANNELS. *J Consult Psychol.* 1967;31(3):248–52.
12. Stillman PL, Sawyer WD. A new program to enhance the teaching and assessment of clinical skills in the People’s Republic of China. *Acad Med.* 1992;67(8):495–9.
13. Liu T, Luo J, He H, Zheng J, Zhao J, Li K. History-taking instruction for baccalaureate nursing students by virtual patient training: A retrospective study. *Nurse Educ Today.* 2018;71:97–104.
14. Aranda JH, Monks SM. Roles and Responsibilities of the Standardized Patient Director in Medical Simulation. *StatPearls. Treasure Island (FL): StatPearls Publishing LLC; 2023.*
15. Zhang S, Soreide KK, Kelling SE, Bostwick JR. Quality assurance processes for standardized patient programs. *Curr Pharm Teach Learn.* 2018;10(4):523–8.

16. Du J, Zhu X, Wang J, Zheng J, Zhang X, Wang Z, Li K. History-taking level and its influencing factors among nursing undergraduates based on the virtual standardized patient testing results: cross sectional study. *Nurse Educ Today*. 2022;111:105312.
17. Edelstein RA, Reid HM, Usatine R, Wilkes MS. A comparative study of measures to evaluate medical students' performance. *Acad Med*. 2000;75(8):825–33.
18. Guagnano MT, Merlitti D, Manigrasso MR, Pace-Palitti V, Sensi S. New medical licensing examination using computer-based case simulations and standardized patients. *Acad Med*. 2002;77(1):87–90.
19. Hawkins R, MacKrell Gaglione M, LaDuca T, Leung C, Sample L, Gliva-McConvey G, Liston W, De Champlain A, Ciccone A. Assessment of patient management skills and clinical skills of practising doctors using computer-based case simulations and standardised patients. *Med Educ*. 2004;38(9):958–68.
20. Maicher KR, Stiff A, Scholl M, White M, Fosler-Lussier E, Schuler W, Serai P, Sunder V, Forrestal H, Mendella L, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. *Med Teach*. 2023;45(3):279–85.
21. Hauze SW, Hoyt HH, Frazee JP, Greiner PA, Marshall JM. Enhancing nursing education through affordable and realistic holographic mixed reality: the virtual standardized patient for clinical simulation. *Adv Exp Med Biol*. 2019;1120:1–13.
22. Kelly S, Smyth E, Murphy P, Pawlikowska T. A scoping review: virtual patients for communication skills in medical undergraduates. *BMC Med Educ*. 2022;22(1):429.
23. Maicher KR, Zimmerman L, Wilcox B, Liston B, Cronau H, Macerollo A, Jin L, Jaffe E, White M, Fosler-Lussier E, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. *Med Teach*. 2019;41(9):1053–9.
24. Borja-Hart NL, Spivey CA, George CM. Use of virtual patient software to assess student confidence and ability in communication skills and virtual patient impression: A mixed-methods approach. *Curr Pharm Teach Learn*. 2019;11(7):710–8.
25. Quail M, Brundage SB, Spitalnick J, Allen PJ, Beilby J. Student self-reported communication skills, knowledge and confidence across standardised patient, virtual and traditional clinical learning environments. *BMC Med Educ*. 2016;16:73.
26. Campillos-Llanos L, Thomas C, Bilinski É, Neuraz A, Rosset S, Zweigenbaum P. Lessons Learned from the Usability Evaluation of a Simulated Patient Dialogue System. *J Med Syst*. 2021;45(7):69.
27. Sun Y, Shuohuan W, Yukun L, Shikun F, Xuyi C, Han Z, et al. ERNIE: enhanced representation through knowledge integration [Internet]. arXiv [Preprint]. 2019. Available from: <https://arxiv.org/abs/1904.09223>.
28. Huang L, Hu J, Cai Q, Fu G, Bai Z, Liu Y, et al. The performance evaluation of artificial intelligence ERNIE bot in Chinese National Medical Licensing Examination. *Postgrad Med J*. 2024;qgae062.
29. Han H. The application of the mask detection based on automatic machine learning. *Proceedings of SPIE*; 2022, 12287. p. 124–9.
30. Wang T, Wu Y. Design and practice of teaching demonstration system for water quality prediction experiment based on EasyDL. In *Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*; 2023. p. 1369–74.
31. Zhao Y, Wang P. The Prediction and Investigation of Factors in the Adaptability Level of Online Learning Based on AutoML and K-Means Algorithm. In *2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT)*; 2024. p. 1313–9.
32. Zhou Z. Automatic machine learning-based data analysis for video game industry. In *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*; 2022. p. 732–7.
33. Fink MC, Reitmeier V, Stadler M, Siebeck M, Fischer F, Fischer MR. Assessment of diagnostic competences with standardized patients versus virtual patients: experimental study in the context of history taking. *J Med Internet Res*. 2021;23(3):e21196.
34. Bond WF, Lynch TJ, Mischler MJ, Fish JL, McGarvey JS, Taylor JT, Kumar DM, Mou KM, Ebert-Allen RA, Mahale DN, et al. Virtual standardized patient simulation: case development and pilot application to high-value care. *Simul Healthc*. 2019;14(4):241–50.
35. Maicher K, Danforth D, Price A, Zimmerman L, Wilcox B, Liston B, Cronau H, Belknap L, Ledford C, Way D, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. *Simul Healthc*. 2017;12(2):124–31.
36. Kammoun A, Slama R, Tabia H, Ouni T, Abid M. Generative adversarial networks for face generation: a survey. *ACM Comput Surv*. 2022;55:1–37.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.