Original Paper

Association of Preexamination AI Virtual-Patient Trainer Scores With Scores on a Local Single-Station History-Taking Objective Structured Clinical Examination: Observational Cohort Study

Weishan Zhang[1*], MSc; Yang Liu[1*], MM; Yiying Zhu[1], MSc; Xian Lu[2], BSc; Liping Wu[1], Prof Dr; Minghui Yue[1], MPH; Xiaolin Hong[1], BMed; Oudong Xia[3*], Prof Dr, MD; Chujun Shi[1*], Prof Dr

[1]Medical Simulation Center, Shantou University Medical College, Shantou, Guangdong, China

[2]Department of Medical Physics and Informatics, Shantou University Medical College, Shantou, China

[3]Shantou University Medical College, Shantou, China

[*]these authors contributed equally

Corresponding Author:
Chujun Shi, Prof Dr
Medical Simulation Center
Shantou University Medical College
Number 22 Xinling Road
Shantou, Guangdong, 515041
China
Phone: 86 754 88900459
Email: cjshi@stu.edu.cn

Abstract

Background: Large language model–based virtual-patient systems score history-taking practice, but whether scores correspond to independently scored examination performance remains untested. Existing evaluations emphasize internal platform metrics or simulated-patient fidelity, leaving criterion-related evidence limited.

Objective: This study tests whether mean preexamination AI-trainer practice scores were associated with performance on 1 local objective structured clinical examination (OSCE) history-taking station, whether the association depended on the complaint-matched training case, and whether a model including the trainer score retained predictive performance in held-out students.

Methods: At 1 medical school, second-year students received a required 6-case assignment on the AI-Powered Medical History-Taking Training and Evaluation System (AMTES) and completed an end-of-course OSCE. Each student completed 1 station rated by 1 examiner; interrater and cross-station reliability could not be estimated. The 4 OSCE diagnoses differed from the training diagnoses: 3 shared a presenting-complaint category with an AMTES case, whereas 1 OSCE case had no presenting-complaint match. Platform-to-OSCE record linkage was 474 out of 480 (98.8%); 473 students with analyzable preexamination practice formed the analytic sample. The primary predictor was the mean of normalized scores across practice records time-stamped before each student's examination-day cut-off. Analyses used Spearman correlations, OSCE-case fixed-effects regression, and repeated 5-fold cross-validation.

Results: OSCE scores (mean 81.5, SD 6.7) were associated with platform practice scores (Spearman ρ=0.388; P<.001). With OSCE-case fixed effects, each 1SD increase in the platform score corresponded to a 2.83-point increase in the OSCE score (95% CI 2.26-3.39). The association was observed in the AMTES-unmatched low back/flank pain stratum (Spearman ρ=0.370; n=108) and with a leave-matched-case-out score (Spearman ρ=0.398 vs 0.370 for the full mean among the same 339 students). Under a single cohort-wide cut-off at the first examination day, the association attenuated but persisted among the same 469 students (Spearman ρ decreased from 0.385 to 0.346; paired difference −0.040, 95% CI −0.065 to −0.018; regression coefficient decreased from 2.83 to 2.68 points per SD). In repeated 5-fold cross-validation (n=471), the platform-only model achieved Spearman ρ=0.374 and R^2=0.176. Adding the platform score to the 2 prior physical-examination course scores increased cross-validated R^2 from 0.141 to 0.236 (ΔR^2=0.096; 95% CI 0.053-0.142). Adding practice volume and case breadth did not improve performance (ΔR^2=−0.002).

XSL•FO
RenderX

Conclusions: By linking required curricular practice scores to the score on 1 independently scored local OSCE history-taking station, this observational study provides criterion-related evidence for interpreting AMTES scores. The association did not depend on including the complaint-matched training-case score. The scores also added predictive information beyond the 2 prior physical-examination course scores in internal cross-validation. These findings establish neither causal effects of practice nor within-student transfer, and residual confounding by prior ability cannot be excluded. They support further evaluation of AMTES analytics as a candidate low-stakes formative signal, with independent external validation required before operational use.

(JMIR Med Educ 2026;12:e109752) doi: 10.2196/109752

KEYWORDS

artificial intelligence; large language models; virtual patients; medical history taking; objective structured clinical examination; criterion-related validity; educational measurement; learning analytics; medical education

Introduction

Eliciting an accurate clinical history is a foundational diagnostic skill, yet providing sufficient interactive practice and expert feedback at scale is resource-intensive. Large language model (LLM)–based virtual patients offer interactive and repeatable opportunities to practice history-taking [1,2]. More recent systems can also provide immediate, structured automated feedback [2] and rubric-based scoring with evidence-linked feedback [3], potentially expanding access to practice and feedback. These systems build on earlier computer-based virtual-patient research [4,5] and the broader technology-enhanced simulation literature [6]. Early LLM-based studies examined simulated-patient plausibility and usability [1], interrater agreement between automated and human evaluations of history-content coverage [2], and the stability and transparency of automated rubric-based scoring [3]. A recent scoping review similarly found that evaluations largely emphasized user experience, authenticity, and system performance, whereas objective learning outcomes were assessed infrequently [7]. More broadly, evidence that performance demonstrated in virtual-patient environments translates to independently assessed clinical performance remains limited and heterogeneous [5,7].

A platform may generate consistent practice scores, but such consistency provides evidence of reliability and does not, by itself, establish validity [8,9]. Contemporary validity theory defines validity in terms of the extent to which evidence and theory support the proposed interpretations and uses of scores, rather than as an inherent property of an assessment instrument [10-12]. Interpreting automated practice scores as indicators of history-taking performance therefore requires evidence that these scores relate as expected to a relevant external measure [8,12]. Argument-based approaches to validation provide practical guidance for organizing and evaluating such evidence in health professions education [11,13]. An independently administered objective structured clinical examination (OSCE) is a relevant external criterion for a history-taking trainer because OSCEs are widely used as standardized assessments of clinical performance [14-18]. In this study, the criterion is the score recorded at a single, platform-independent, assessor-blinded local OSCE history-taking station (platform independence is defined in the “OSCE Outcome” section). The study does not constitute multiinstitutional external validation.

Learning analytics are increasingly proposed for describing and predicting learner performance and supporting formative action, but their interpretation and use require appropriate validation [19,20]. Automated scoring also has precedents in medical education, including the automated scoring of constructed responses [21] and research on natural language processing–assisted scoring of patient notes in the former US Medical Licensing Examination Step 2 Clinical Skills examination [22]. Prior LLM-based virtual-patient research has also examined learner perceptions and feasibility [23], while early intervention studies have evaluated educational outcomes in OSCE preparation [24,25]. These studies addressed system performance, learner experience, rater agreement, or intervention effectiveness rather than whether automated scores accumulated during routine practice correspond to performance on a separately administered clinical assessment.

Previous studies of the AI-Powered Medical History-Taking Training and Evaluation System (AMTES) examined automated scoring stability and human-AI consistency [3], associations with summative educational outcomes under voluntary use [26], and learner acceptance [27]. However, they did not determine whether AMTES scores generated during routine practice correspond to performance on an independently administered clinical assessment. This study therefore addresses a distinct, noncausal measurement question in the context of required curricular use. To our knowledge, cohort-scale evidence linking automated trainer scores derived from practice sessions submitted before an examination to independently assessed OSCE performance remains scarce.

Whether this association holds matters for potential deployment. If AMTES performance is associated with the score recorded on a separately administered, local, single-station OSCE, platform analytics may provide a scalable, low-stakes formative signal to guide faculty review and targeted educational support. If not, the automated scores may primarily reflect platform-specific performance. Even if an association is observed, it may reflect shared prior ability or conscientious engagement in both settings, or rehearsal matched to the examination content rather than a broader history-taking performance signal. To address these alternative explanations, this study used a local OSCE aligned with established curriculum-blueprinting principles [16-18]. The OSCE was administered independently of AMTES, and examiners had no access to students’ platform records.

Accordingly, this study addressed 3 research questions (RQs):

XSL•FO
RenderX

RQ1 (criterion-related evidence): Are students' mean AMTES practice scores associated with the score recorded at a local, single-station, independently scored OSCE history-taking station?

RQ2 (dependence on the complaint-matched AMTES case): Does the association persist in the AMTES-unmatched presenting-complaint stratum and when the score contribution from the presenting-complaint-matched AMTES case is omitted from the platform score?

RQ3 (internal cross-validated prediction): How well does a model predict held-out students' OSCE scores using prior examination scores and AMTES measures derived from time stamp–restricted practice records, without assigned-case information?

Methods

Study Design

This observational study drew on the complete eligible curricular cohort within a course that included 2 required components: a 6-case AMTES assignment and an end-of-course OSCE. Practice records retained for construction of the primary predictor were time-stamped between May 10 and June 26, 2026. The OSCE was administered over 5 days (June 22-26, 2026), with different classes examined on different days. As the practice and examination periods overlapped at the cohort level, only practices time-stamped before each student's own OSCE cut-off, as defined below, were included in the primary predictor. Neither AMTES exposure nor practice dose was randomized. Study size was determined by the eligible curricular cohort rather than by an a priori power calculation. Reporting followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [28]. Supplementary numeric results are provided in Multimedia Appendix 1, and extended methods and results are provided in Multimedia Appendix 2.

Setting, Participants, and the AI Trainer

Eligibility for the analytic cohort was defined by the curricular and assessment arrangements established before examination performance was observed. Eligible participants were second-year students in the 2024-entry Clinical Medicine program at Shantou University Medical College whose curriculum included both the 6-case AMTES assignment and the concurrently scheduled end-of-course OSCE. Some students taking the same history-taking course completed the AMTES assignment but were not scheduled to take the OSCE under their curricular programs' assessment arrangements; they were therefore ineligible for this analysis. One repeat-year student who returned only to resit the OSCE without participating in the corresponding AMTES course component was also ineligible. Eligibility was not determined by actual AMTES engagement, successful record linkage, or OSCE performance. Accordingly, whether a student had a retained scored record for each of the 6 assigned cases was treated as an exposure characteristic rather than as an eligibility or analytic-inclusion criterion.

AMTES, the LLM-based virtual-patient trainer described previously [3], was deployed as a required self-study in the students' first formal history-taking course. The assignment comprised 6 history-taking cases involving cough, fever, urinary frequency, chest pain, headache, and abdominal pain. Completing each assigned case once counted toward the course requirement; repeat practice was optional and earned no additional credit. After each practice session submitted for evaluation, the platform generated item-level rubric feedback, overall feedback, and an overall score. For the primary predictor, retained platform records were limited to evaluated practice records with valid scores that were time-stamped before the student-specific cut-off defined below.

AI Trainer Configuration During the Study Period

Virtual-patient dialogue was generated by DeepSeek-V4-flash (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co Ltd). Practice scoring was fully automated and comprised item-level scoring by DeepSeek-V3.2 at temperature 0 using a shared scoring prompt template, followed by automated postprocessing of the item scores before the resulting score was returned to the student. The scores analyzed here are those received by students. The configured model identifiers and platform-level prompt templates were the same for all students and shared across the 6 training cases; each case supplied its own clinical content and scoring rubric. According to the platform configuration and deployment records, these model identifiers and templates remained unchanged during the study period, and AMTES version 4.0 was not redeployed. The AMTES workflow, multilevel verification procedures, and comprehensive history-taking scoring prompt are described in our previous system-development report and its multimedia appendices 1-3 [3]. The model configuration used for the present deployment is specified above.

OSCE Outcome

At the end of the course, students completed an OSCE history-taking station involving 1 of 4 presenting complaints: chest pain, cough/dyspnea, fever/rash, or low back/flank pain. Each student completed 1 station and received a single rating from the examiner assigned to that station; no performance was independently double-scored. Within each examination session, all 4 cases operated concurrently as parallel stations. Examination staff acting as guides directed waiting students to available stations without using student identity, class, prior performance, or preference to determine station assignment. No predetermined candidate-to-station sequence, formal random-allocation procedure, or assignment log was used. We therefore describe the process as availability-driven allocation by examination staff rather than as randomization.

Examiners were blinded specifically to AMTES data: they were neither provided with nor given access to individual students' AMTES practice records or platform-derived scores before or during scoring. As used in this paper, platform independence means that a different instrument was administered and scored outside the platform by assessors masked to platform data. It does not imply construct independence, because both measures target history-taking; statistical independence of the 2 scores; separation of personnel, as several authors developed or

XSL•FO
RenderX

deployed AMTES (conflicts of interest); or any claim about the reliability or generalizability of the OSCE score itself. The examination followed a structured OSCE format [14-16] and used standardized patients (SPs). Each local station consisted of a 20-minute, 1-on-1 history-taking consultation. Because of available examiner and SP staffing, assessment of the full eligible cohort was distributed across 5 days using parallel, case-specific stations. The local scoring rubric was developed with reference to China's National Medical Licensing Examination. It comprised case-specific history-content items and a 6D communication-skills and comprehensive-performance block (hereafter referred to as the *communication block*). The official recorded station score, used in all analyses, ranged from 0 to 100 points and comprised 70 points for history content and 30 points for the communication block. The total was calculated as the sum of the corresponding item-level scores. Before the examination, all SPs and examiners received common training intended to promote consistency in case portrayal and scoring. Such training and standardization are recognized quality-assurance safeguards for performance-based assessments [29].

Nine classes sat the examination, 1 per half-day session (4 morning and 5 afternoon) over 5 days; class and session therefore coincided, so the class-grouped cross-validation reported below is equivalently session-grouped. Sessions used 4-7 stations, depending on class size; when a case required more than 1 station in a session, each station was staffed by a different SP and scored by a different examiner. Thirteen SPs took part, each portraying 1 case (4 SPs portrayed the chest pain case, and 3 SPs portrayed each of the other cases) and participating in 4 or 5 of the 9 sessions. Twelve examiners scored the history-taking stations; they were assigned to station slots rather than to cases and therefore scored different cases across sessions. Examiner identity was not recorded with individual scores, so examiner-specific effects could not be estimated. The station deployment grid is provided in Table S7 in Multimedia Appendix 1, with further administration details provided in Multimedia Appendix 2.

The 4 OSCE cases involved diagnoses that were not represented among the 6 AMTES training cases: gastroesophageal reflux disease presenting with chest pain, pleural effusion presenting with cough/dyspnea, dengue fever presenting with fever/rash, and renal stone presenting with low back/flank pain. For the content-overlap analyses, 3 OSCE cases were matched to an AMTES case solely based on a prespecified presenting-complaint category: chest pain with the myocardial infarction training case, cough with the pneumonia training case, and fever with the liver abscess training case. The diagnoses differed within all 3 matched pairs. The renal-stone OSCE case, designated as low back/flank pain, had no corresponding case among the 6 assigned AMTES cases with the same prespecified presenting-complaint category and was therefore classified as the AMTES-unmatched presenting-complaint case. In this article, "AMTES unmatched" refers only to the absence of a presenting-complaint match among the assigned AMTES cases and does not imply that students had no prior exposure to or practice with that complaint in other settings. Within the matched pairs, the OSCE and AMTES case scripts also differed in features such as onset, precipitating factors, and illness evolution. Nevertheless, generic symptom-elicitation domains and complaint-level interviewing schemas could still overlap (Table S3 in Multimedia Appendix 1).

Linkage and Measures

OSCE records were linked to platform records using a crosswalk between student numbers and platform user IDs. Inclusion in the analytic sample required an OSCE history-taking score and at least one retained AMTES practice record time-stamped before the examination (Figure 1); this analytic requirement was distinct from curricular eligibility.

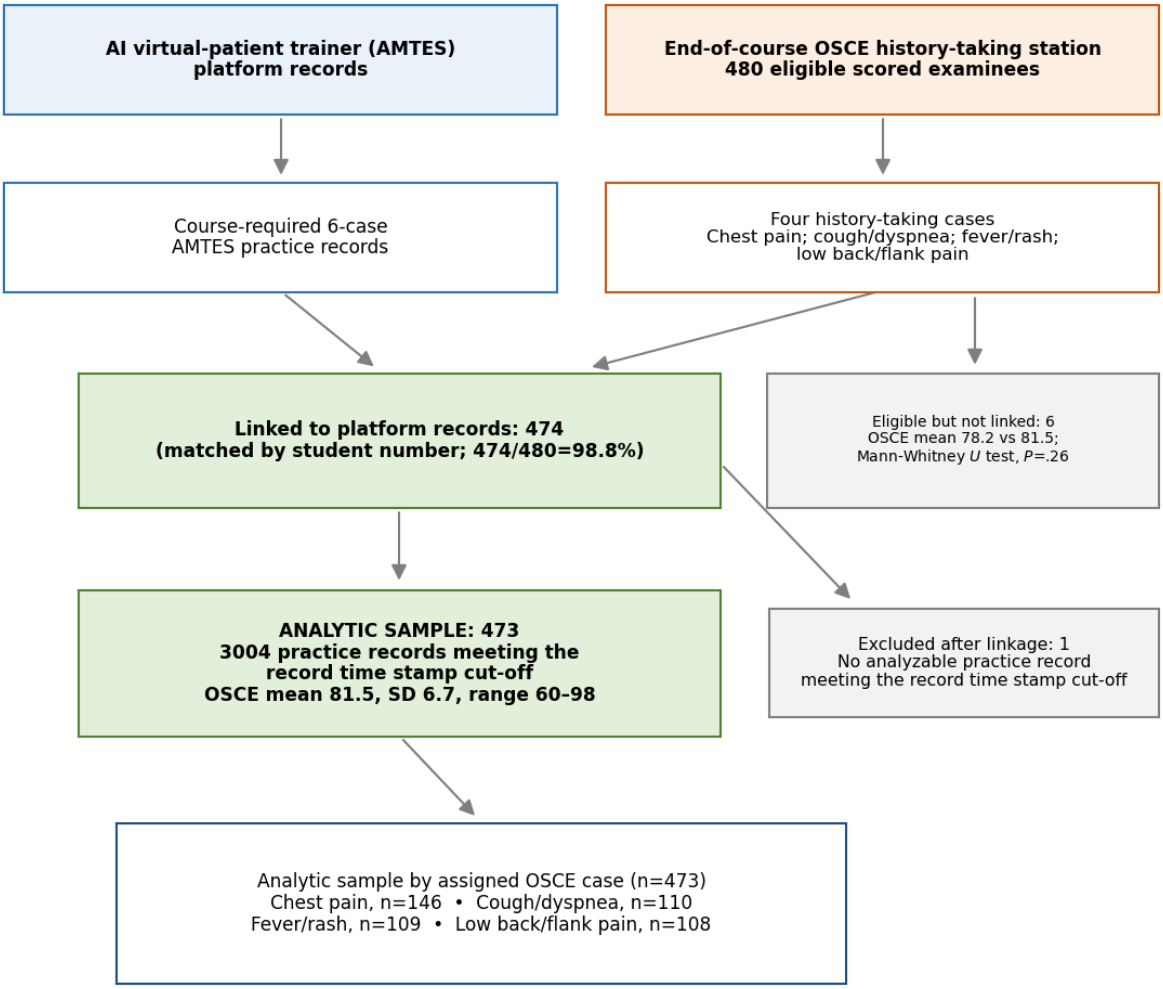
The primary predictor was each student's mean normalized score across retained practices, chosen to represent the student's average score across recorded practice encounters, with each encounter contributing equally. As this approach gives greater weight to cases practiced repeatedly, a case-balanced mean (calculated by averaging scores within each practiced case and then across practiced cases) was evaluated as a sensitivity analysis (see the "Results" section; also see Table S1 in Multimedia Appendix 1).

Each retained practice score was divided by the maximum score for its case, placing scores from cases with different rubric maxima on a common scale from 0 to 1. The case-specific maximum was 62 points for the cough case and 70 points for each of the other 5 AMTES cases. For descriptive reporting, cough scores were also expressed on an equivalent 70-point scale; this linear rescaling did not change their normalized values. Practice-record time stamps indicate when students submitted completed practices for evaluation. The normalized practice scores were then averaged using only records time-stamped before a conservative examination-day cut-off, defined as the earliest scheduled OSCE start time (8 AM) on each student's examination date. All analyzed practice scores are traceable to their original scoring records.

Secondary predictors included the leave-matched-case-out score, practice volume, and case breadth. Practice volume was defined as the number of retained practices, and case breadth as the number of distinct AMTES cases practiced. The leave-matched-case-out score was calculated by excluding practices from the AMTES case matched to the student's assigned OSCE case by presenting-complaint category and then averaging the remaining normalized practice scores. This score was defined only for students who had practiced both the complaint-matched AMTES case and at least one additional case. The measured baseline covariates were the 2 examinations from the preceding physical-examination course, which was a prerequisite for the history-taking course: its written theory examination (Clinical Basic Skills I; 0-100) and its practical OSCE examination (total score). Both examinations were taken in the preceding semester, before the practice window. Both assess physical examination rather than history-taking and are the only prior physical-examination course scores available for this cohort. They were used as baseline covariates in adjusted association models and as predictors in models without assigned OSCE-case information.

XSL•FO
RenderX

Figure 1. Participant flow and linkage of eligible objective structured clinical examination (OSCE) examinees to preexamination platform records (analytic sample, N=473). AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



Student sex was obtained from the examination score sheets. All students belonged to the same curricular year group and took the course together. Date of birth was not available in the examination or platform records used to assemble the analytic dataset; age was therefore not reported.

The primary outcome was the official recorded OSCE station total, comprising history content and communication/comprehensive performance. We examined whether the content-focused AMTES practice score was associated with this broader local assessment of a history-taking encounter, without assuming that the 2 measures represented identical constructs. The history-content subscore, which more directly aligned with the AMTES rubric, was analyzed as a supporting outcome; the communication block provided a comparison with a less directly aligned component. The official total, unchanged, is the score used for progression, resit, and transcript decisions.

Statistical Analysis

Overview

Analyses first addressed the RQs, followed by sensitivity analyses, robustness checks, and additional design checks.

Primary Association (RQ1)

Spearman correlations quantified bivariate associations. Ordinary least squares regression with OSCE-case fixed effects was used to estimate semistandardized slopes. Predictors were *z* scored, whereas the outcome was retained on its original point scale. The fixed effects were included to account for average score differences across the 4 OSCE cases. They adjust only for between-case mean differences and do not, by themselves, account for examiner or station measurement error, which is addressed as a limitation below. For descriptive visualization, students were grouped by quartiles of the platform score, and OSCE scores were compared across quartiles using the Kruskal-Wallis test.

XSL•FO
RenderX

Dependence on the Complaint-Matched AMTES Case (RQ2)

Dependence was examined in 2 ways: (1) by estimating the association within the AMTES-unmatched low back/flank pain OSCE stratum and (2) by examining the association between the leave-matched-case-out platform score and OSCE performance, with the complaint-matched AMTES case score reported as a descriptive comparison. The leave-matched-case-out analysis was restricted to students who had practiced the complaint-matched AMTES case and at least one additional case. All comparisons between the leave-matched-case-out score and the overall mean were computed for the same students.

Internal Cross-Validated Prediction (RQ3)

Candidate predictors were the 2 prior physical-examination course scores and AMTES measures derived from records meeting the student-specific time stamp cut-off. The full model without assigned OSCE-case information included both prior examination scores, the platform score, practice volume, and case breadth. Out-of-sample performance was assessed using internal 5-fold cross-validation, repeated 200 times with a fixed seed. Model fitting and predictor standardization were performed within the training folds. The evaluated models included the platform score alone, both prior examinations alone, their combination, and the full model; OSCE-case fixed effects alone and in combination with the platform score were evaluated as conditional comparators. Cross-validated Spearman ρ, R^2, and mean absolute error (MAE) were calculated from pooled out-of-fold predictions within each repetition and averaged across repetitions. Models incorporating the baseline covariates were fitted in the subset of 471 students with data available for both examination scores. Bootstrap 95% CIs were calculated using 2000 paired student resamples. In each resample, performance was calculated separately from the out-of-fold predictions for each repetition and then averaged across repetitions, using the same resampled students for all model contrasts. Models and folds were held fixed; therefore, these intervals are conditional on the existing out-of-fold predictions and do not incorporate additional uncertainty from refitting, fold generation, or examination clustering (Multimedia Appendix 2).

Sensitivity, Robustness, and Post Hoc Analyses

Sensitivity and robustness checks addressed outcome definition; construct alignment, using the communication block as a less directly aligned comparison outcome; practice dose under fixed exposure; the analysis window; measured baseline physical-examination performance; case breadth; predictor weighting; within-case outcome scaling; case-assignment balance; and class- and examination-day–grouped validation. Post hoc checks explicitly addressed the influence of low scores; score construction and practice strategy; examination-day and case-by-day effects; recorded station-slot labels; and the platform score's incremental contribution beyond measured baseline and practice covariates. Full specifications are provided in Multimedia Appendix 2.

Additional Analyses

Additional analyses comprised a cohort-wide preexamination cut-off; a comparison of students included in and excluded from the leave-matched-case-out analysis; a common-scale comparison of the history-content and communication components; and sex-related analyses of differential performance, differential prediction, and out-of-fold subgroup calibration. The cohort-wide cut-off included only practices time-stamped before the first examination day (June 22, 2026). The primary association and cross-validation comparisons were repeated in the same 469 students, whereas RQ2 analyses were repeated in their respective defined-score samples and reported descriptively. The 2 dependent Spearman correlations were compared using a paired student bootstrap. Exploratory within-student discontinuity and practice-pattern checks were also conducted. All analyses are specified in Multimedia Appendix 2.

Analysis Population and Missing Data

No values were imputed; each model used records with complete data for its variables. Baseline-adjusted analyses were restricted to complete cases with both prior physical-examination course scores, and the unadjusted model was refitted in the same subset. OSCE scores of linked and unlinked eligible examinees were compared using the Mann-Whitney *U* test.

RQ1 was the single primary association, whereas RQ2 and RQ3 were supportive research-question analyses. Per-case, quartile, component, sensitivity, robustness, post hoc, and additional analyses (including the sex-related and cohort-wide-cut-off analyses) were not treated as independent confirmatory tests; therefore, no multiplicity adjustment was applied. Precision was summarized using 2-sided 95% CIs throughout. Regression (ordinary least squares) CIs were calculated using heteroskedasticity-consistent SEs (type 3; HC3), and CIs for Spearman correlations were calculated using the Fisher *z* transformation.

Analyses were conducted using Python 3.12 (pandas, SciPy, and statsmodels; Python Foundation). Analysis and figure-generation code and derived data are available as described in the "Data Availability" section.

OSCE Measurement Evidence and Scoring Sensitivity

The OSCE stations were aligned with the local course blueprint and used predefined case-specific scoring criteria, consistent with established OSCE blueprinting principles [16-18]. Assessor masking and common pre-OSCE training are described above; however, no formal cross-examiner calibration or independent double-rating data were available for this administration. Case-specific Cronbach α was calculated for the 6D communication block. For the history-content checklist, α was reported descriptively rather than treated as sufficient or primary evidence of score reliability, because internal consistency does not capture important rater- and station-related sources of measurement error in performance assessment [9]. In a scoring-rule sensitivity analysis, history-content items were dichotomized using full- and any-credit rules, and the platform-content association was recalculated. As no performance was independently rated by a second examiner

XSL•FO
RenderX

and each student completed only 1 station, interrater reliability and student-level cross-station generalizability could not be estimated in this study [9,30]. Detailed results are reported in Table S2 in Multimedia Appendix 1.

Ethical Considerations

This study received ethical approval from the Ethics Committee of Shantou University Medical College (SUMC-2026-017) and was conducted in accordance with the Declaration of Helsinki. The 6-case AMTES assignment was a required course component and contributed to coursework credit; students received no additional compensation for the research use of their records. Written informed consent covered the research use of platform learning records and their linkage to examination records, including OSCE scores. Before analysis, the linked data were pseudonymized and stored on password-protected institutional servers accessible only to authorized members of the research team.

Results

Sample, Linkage, and Participant Flow

Of 480 eligible examinees with scored OSCEs, 474 (98.8%) were linked to platform records. The analytic sample comprised 473 students with an OSCE history-taking score and analyzable preexamination practice records; 1 linked examinee had no analyzable preexamination practice record (Figure 1). These students contributed 3004 retained preexamination practices, with a median of 6 (IQR 6-7) practices across a median of 6 distinct cases (IQR 5-6 distinct cases; Table 1). Although 366 (77.4%) students had at least six retained practice records, 349 (73.8%) had at least one retained, validly scored record for each of the 6 assigned cases. Students who had not completed the assignment were retained because completion was treated as an exposure characteristic rather than an inclusion criterion (Table 1; Table S1 in Multimedia Appendix 1). OSCE scores had a mean of 81.5 (SD 6.7), a median of 82, and a range of 60-98. Case counts were 146 for chest pain, 110 for cough/dyspnea, 109 for fever/rash, and 108 for low back/flank pain (Table 1).

Mean OSCE scores differed across the 4 allocated case groups (case means 77.1-84.1; 1-way $F_{3,469}$=31.06; P<.001; case-only R^2=0.166). These between-case mean differences cannot be attributed uniquely to case difficulty; the OSCE-case fixed effects used in subsequent analyses adjust for the composite differences without decomposing them. All 9 classes were represented in every case. Students allocated to the 4 cases had similar mean AMTES practice scores (mean normalized scores 0.738, 0.729, 0.730, and 0.727; $F_{3,469}$=0.38; P=.77; R^2=0.002) and similar scores on both prior physical-examination course examinations (theory: case-group means 67.5-68.7; $F_{3,468}$=0.22; P=.88; practical: 36.9-37.5; $F_{3,468}$=0.29; P=.83; n=472 for each). These measured similarities provide no evidence of systematic assignment based on the observed student characteristics, but they do not establish formal randomization or balance in unmeasured characteristics.

Among the 6 unlinked eligible examinees, the mean OSCE score was 78.2, compared with 81.5 among the 474 linked eligible examinees (Mann-Whitney U test, P=.26). Given the small number of unlinked examinees, this comparison is imprecise.

XSL•FO
RenderX

Table 1. Participant, OSCE[a], and preexamination AMTES[b] practice characteristics (N=473).

Characteristics	Values
Participants, n	
Analytic sample	473
Sex, n (%)	
Male	307 (64.9)
Female	166 (35.1)
OSCE performance	
Total score, mean (SD)	81.5 (6.7)
Total score, median (range)	82 (60-98)
Assigned OSCE case, n	
Chest pain	146
Cough/dyspnea	110
Fever/rash	109
Low back/flank pain[c]	108
OSCE score, mean (SD)	
Chest pain	84.1 (6.3)
Cough/dyspnea	77.1 (5.7)
Fever/rash	80.9 (6.4)
Low back/flank pain[c]	83.2 (6.0)
Preexamination AMTES practice	
Mean normalized score, mean (SD)	0.73 (0.09)
Retained practices, median (IQR)	6 (6-7)
Distinct cases practiced, median (IQR)	6 (5-6)
Students with ≥6 retained practice records, n (%)	366 (77.4)
Students who practiced all 6 distinct assigned cases, n (%)	349 (73.8)

[a]OSCE: objective structured clinical examination.

[b]AMTES: AI-Powered Medical History-Taking Training and Evaluation System.

[c]Low back/flank pain was the AMTES-unmatched OSCE presenting-complaint stratum (ie, no assigned AMTES case shared its prespecified presenting-complaint category).

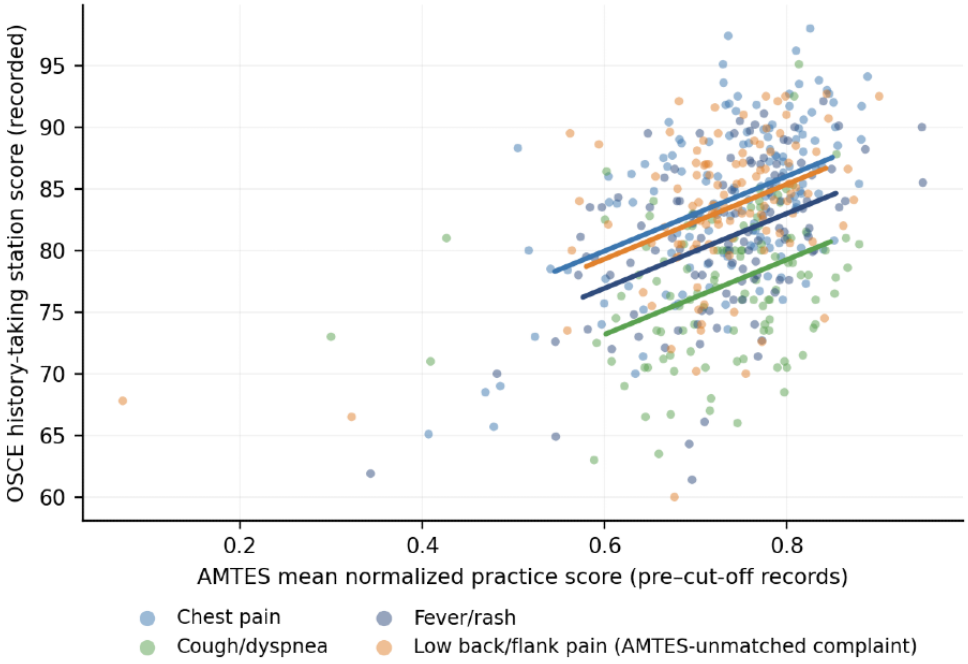
RQ1: AI Trainer Performance Is Associated With the Recorded OSCE Station Score

Mean platform practice performance was associated with the recorded OSCE history-taking station score (Spearman ρ=0.388; *P*<.001; N=473; Figure 2). A similar correlation was observed after OSCE scores were standardized within case (Spearman ρ=0.393; *P*<.001; Table S1 in Multimedia Appendix 1). In the OSCE-case-fixed-effects model, each 1SD increase in platform score was associated with a 2.83-point increase in OSCE score (95% CI 2.26-3.39; *P*<.001; Table 2). Model R^2 increased from 0.166 with case fixed effects alone to 0.344 after the platform score was added.

This estimate was equivalent to approximately 0.42 of the observed OSCE score SD (6.7 points). This was a between-student association; it does not imply that additional practice would raise an individual student's OSCE score by 2.83 points. Mean OSCE scores increased across ascending quartiles of platform performance, from 77.9 in Q1 to 81.1, 82.4, and 84.8 in Q2-Q4, respectively (Figure 3). OSCE score distributions differed across the 4 quartiles (Kruskal-Wallis *H*=60.46, *df*=3; *P*<.001). Practice volume and case breadth showed smaller bivariate correlations with OSCE score (Spearman ρ=0.213 and Spearman ρ=0.169, respectively; both *P*<.001). The case-balanced score yielded a Spearman ρ of 0.399 (95% CI 0.32-0.47) and an association of +2.92 OSCE points per SD (95% CI 2.34-3.49), compared with Spearman ρ=0.388 and +2.83 points for the practice-weighted score (Table S1 in Multimedia Appendix 1). Among the 471 students with both prior physical-examination course scores, adjustment for the 2 examinations reduced the platform coefficient in the case-fixed-effects model from +2.80 (95% CI 2.23-3.36) to +2.15 OSCE points per SD (95% CI 1.61-2.69; *P*<.001). A post hoc nested model that also included practice volume and case breadth yielded +2.16 points per SD (95% CI 1.60-2.72; *P*<.001; Table 2).

XSL•FO
RenderX

Figure 2. Mean normalized AI-trainer practice score versus recorded objective structured clinical examination (OSCE) history-taking station score (N=473). AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



XSL•FO
RenderX

Table 2. Primary and supportive analyses of the association between AMTES[a] performance and the recorded OSCE[b] history-taking station score.

Analysis	n	Statistic	Estimate	95% CI	*P* value
Primary association					
Platform score and OSCE station total	473	Spearman ρ	0.388	0.31-0.46	<.001
OSCE-case fixed effects only	473	Model R^2	0.166	N/A[c]	N/A
OSCE-case fixed effects plus platform score	473	B/OSCE points per SD[d]	2.83	2.26-3.39	<.001
Model fit after adding the platform score	473	Model R^2	0.344	N/A	N/A
Cohort-wide preexamination cut-off sensitivity analysis[e]					
Platform score and OSCE station total, primary window → June 22, 2026, cut-off	469	Spearman ρ	0.385 → 0.346	difference −0.040 (−0.065 to −0.018)	N/A
OSCE-case fixed effects plus platform score, primary window → June 22, 2026, cut-off	469	B/OSCE points per SD[d]	2.83 → 2.68	2.11-3.26 (cut-off)	<.001
Dependence on the complaint-matched AMTES case					
AMTES-unmatched low back/flank pain stratum[f]	108	Spearman ρ	0.370	0.19-0.52	<.001
Leave-matched-case-out platform score[g]	339	Spearman ρ	0.398	0.30-0.48	<.001
Overall mean score in the same 339 students (same-sample comparator)	339	Spearman ρ	0.370	0.27-0.46	N/A
Internal repeated cross-validation[h]					
Platform score only	471	Spearman ρ, R^2, and MAE[i]	0.374/0.176/4.85	N/A	N/A
Full model without assigned OSCE-case information (platform score + 2 prior examinations + practice volume + case breadth)	471	Spearman ρ, R^2, and MAE	0.437/0.235/4.71	0.36-0.51/0.15-0.30/4.40-5.04	N/A
Calibration of the full model without assigned OSCE-case information	471	Calibration-in-the-large and slope	0.000 and 0.949	slope 0.80-1.10	N/A
Baseline and covariate adjustment					
Platform score, unadjusted (complete-case subset)	471	B/OSCE points per SD[d]	2.80	2.23-3.36	<.001
Platform score, adjusted for both prior physical-examination course scores	471	B/OSCE points per SD[d]	2.15	1.61-2.69	<.001
Platform score, adjusted for both examinations, practice volume, and case breadth	471	B/OSCE points per SD[d]	2.16	1.60-2.72	<.001
Component outcomes					
Difference in fully standardized β, history-content subscore minus communication block	473	Δβ	0.202	0.091-0.314	<.001

[a]AMTES: AI-Powered Medical History-Taking Training and Evaluation System.

[b]OSCE: objective structured clinical examination.

[c]N/A: not applicable.

[d]B coefficients represent the difference in OSCE score associated with a 1SD higher platform score. Models included OSCE-case fixed effects and used heteroskedasticity-consistent SEs (type 3) to calculate 95% CIs.

[e]This sensitivity analysis restricted practices to those time-stamped before June 22, 2026, and compared this cut-off with the primary window in the same 469 students. The paired difference in Spearman ρ was computed from unrounded correlations. Full results are in Table S4 in Multimedia Appendix 1.

[f]Low back/flank pain was the AMTES-unmatched OSCE stratum; no assigned AMTES case shared its prespecified presenting-complaint category.

[g]The leave-matched-case-out score excluded practices from the AMTES case matching the assigned OSCE presenting complaint; the comparator used

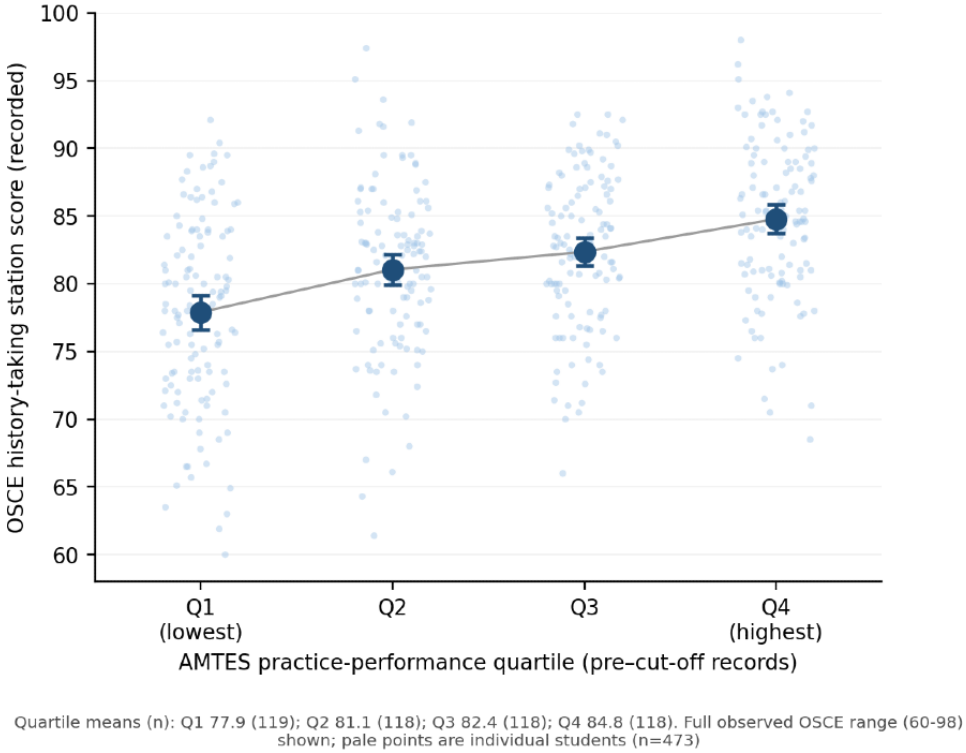
XSL•FO
RenderX

the overall mean score in the same students.

[h]Repeated 5-fold cross-validation used within-fold standardization and model fitting. Reported metrics are averages across repetitions from held-out predictions; models using prior examinations included 471 complete cases. Additional models are in Table S5 in Multimedia Appendix 1.

[i]MAE: mean absolute error.

Figure 3. Unadjusted objective structured clinical examination (OSCE) history-taking scores by quartile of AI-trainer practice performance. AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



Component Outcomes: History-Content Subscore and Communication Block

With OSCE-case fixed effects, the platform score was associated with the history-content subscore at +2.30 points per SD (95% CI 1.83-2.78) and with the communication block at +0.53 points per SD (95% CI 0.27-0.78). The association with the communication block was weaker but not zero. On a common scale, the fully standardized coefficients were 0.391 (95% CI 0.310-0.472) for content and 0.189 (95% CI 0.098-0.279) for communication, a difference of 0.202 (95% CI 0.091-0.314; *P*<.001; Table 2). Standardization equates scale but not reliability. Moreover, the communication block is ceiling-prone and nearly uncorrelated with content (Spearman ρ=0.032). Thus, this difference reflects alignment with observed scores rather than relationships between true scores; no disattenuated estimates are reported. The association was also present when the content subscore replaced the station total as the criterion (Spearman ρ=0.370; 95% CI 0.289-0.445). The difference between the correlations for the total score and content subscore was small (Spearman Δρ=0.018; 95% CI −0.018 to 0.057), although the 2 outcomes shared 70 of the 100 station points (see Table S1 and Figure S1 in Multimedia Appendix 1).

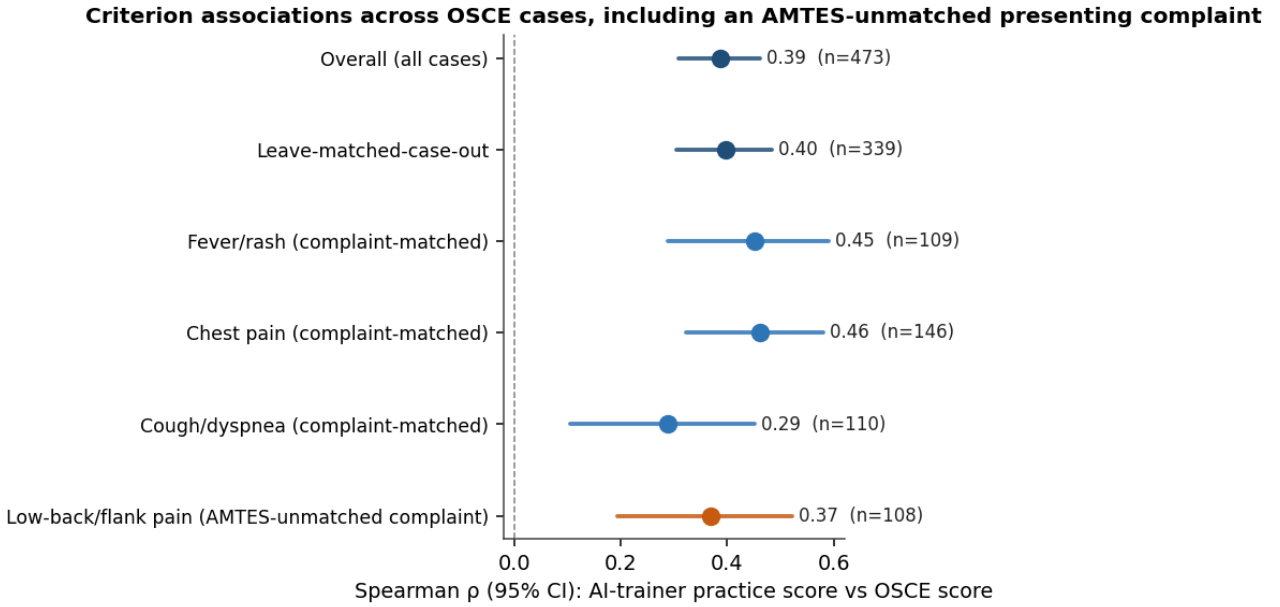
RQ2: The Association Does Not Depend on Including the Complaint-Matched AMTES-Case Score

Within the low back/flank pain OSCE stratum, whose presenting complaint had no match among the assigned AMTES cases, platform performance was associated with OSCE score (Spearman ρ=0.370; 95% CI 0.19-0.52; *P*<.001; n=108). Positive associations were also observed in the 3 presenting-complaint-matched strata: chest pain (Spearman ρ=0.462; n=146), cough/dyspnea (Spearman ρ=0.288; n=110), and fever/rash (Spearman ρ=0.452; n=109; Figure 4 and Table S1 in Multimedia Appendix 1).

Among the 341 students who had practiced the training case matching their OSCE presenting complaint, 339 had practiced at least one additional case and therefore had a defined leave-matched-case-out score. In these 339 students, the leave-matched-case-out score showed a numerically higher correlation with the OSCE station score than the overall mean score (Spearman ρ=0.398; 95% CI 0.30-0.48 vs Spearman ρ=0.370; 95% CI 0.27-0.46). The complaint-matched training-case score itself was less strongly correlated with the OSCE score (Spearman ρ=0.174; *P*=.001; n=341; descriptive comparison). However, the leave-matched-case-out score remained highly correlated with the full platform aggregate (Spearman ρ=0.917; n=339), indicating that this analysis primarily assesses whether the overall association depends on the matched case rather than providing independent evidence of generalization. Both analyses involve between-student comparisons and do not establish within-student transfer across clinical presentations.

XSL•FO
RenderX

Figure 4. Associations between AI-trainer scores and recorded objective structured clinical examination (OSCE) station scores overall, for leave-matched-case-out scores, and by assigned OSCE case. AMTES: AI-Powered Medical History-Taking Training and Evaluation System.



The 26 students in the complaint-matched strata without a defined leave-out score had fewer practices (median 2.5 vs 6), and lower platform scores (0.654 vs 0.739) than the 339 included students. Within the 365-student complaint-matched stratum, the platform-score-by-inclusion interaction was −0.70 points per SD (95% CI −2.96 to 1.55; *P*=.54), providing no evidence of a detectable difference in slope. However, these comparisons do not establish the direction or magnitude of any selection effect (see Table S6 in Multimedia Appendix 1; also see Multimedia Appendix 2).

RQ3: Internal Cross-Validated Performance Without Assigned OSCE-Case Information

Across 200 repetitions of 5-fold cross-validation among the 471 students with both prior examination scores, the platform score alone yielded held-out predictive performance for OSCE scores of Spearman ρ=0.374, R^2=0.176, and an MAE of 4.85 points. The full model without assigned OSCE-case information yielded Spearman ρ=0.437 (bootstrap 95% CI 0.36-0.51), R^2=0.235 (95% CI 0.15-0.30), and an MAE of 4.71 points (95% CI 4.40-5.04), with a calibration slope of 0.949 (95% CI 0.80-1.10).

Adding the platform score to the 2 prior examinations increased cross-validated R^2 from 0.141 to 0.236 (ΔR^2=0.096; 95% CI 0.053-0.142; computed at full precision). By contrast, adding practice volume and case breadth did not improve mean cross-validated performance in this cohort (ΔR^2=−0.002; 95% CI −0.012 to 0.008). Additional model comparisons, calibration details, and prediction-error summaries are reported in Table S5 in Multimedia Appendix 1. Metrics from models that also included the assigned OSCE case are conditional on that assignment (see Figures S2 and S3 in Multimedia Appendix 1; also see Multimedia Appendix 2).

Examination-Week Overlap: Cohort-Wide Preexamination Cut-Off

Applying a single cohort-wide cut-off before the first examination day (June 22, 2026) to all students excluded 391 of the 3004 (13.02%) practices and left 469 students with an available predictor. In these 469 students, the association attenuated modestly: Spearman ρ decreased from 0.385 to 0.346 (95% CI 0.26-0.42; paired difference −0.040, 95% CI −0.065 to −0.018), and the case-fixed-effects coefficient decreased from +2.83 to +2.68 points per SD (95% CI 2.11-3.26), while the quartile gradient remained monotonic. Both RQ2 analyses remained positive (AMTES-unmatched stratum: Spearman ρ=0.323; n=107; leave-matched-case-out analysis: Spearman ρ=0.375; n=321). The cut-off analysis does not identify the cause of the attenuation or exclude the possibility of unrecorded exchanges of examination information, which would leave no practice record. Exploratory timing and practice-pattern analyses are reported in Table S4 in Multimedia Appendix 1 and in Multimedia Appendix 2.

Sex-Related Differences in Scores and Prediction

Sex was recorded for all 473 students (307 male and 166 female). Women scored higher than men on the OSCE (83.1 vs 80.7 points; Cohen *d*=0.36; *P*<.001) and on the platform predictor (0.76 vs 0.72; Cohen *d*=0.50; *P*<.001). The association was present in both sexes (men: Spearman ρ=0.39; +2.65 OSCE points per SD, 95% CI 1.99-3.31; women: Spearman ρ=0.29; +2.84 points per SD, 95% CI 1.92-3.77). No detectable difference in slope was found (interaction +0.30 points per SD, 95% CI −0.88 to 1.49; *P*=.61), although the CI was wide. Adjusting the primary model for sex changed the platform coefficient from +2.83 to +2.71. At the same platform score and assigned OSCE case, women had a higher conditional mean OSCE score (+1.10 points, 95% CI 0.12-2.09; *P*=.03). This difference was concentrated in the history-content component

XSL•FO
RenderX

(+1.03 points, 95% CI 0.20-1.85), whereas the difference in the communication block was small (+0.07 points, 95% CI −0.41 to 0.56). Out-of-fold mean prediction errors from the full model without assigned OSCE-case information were +0.25 points in women and −0.13 points in men; the corresponding calibration slopes were 0.95 and 0.92. Adding sex did not improve prediction (ΔR^2=−0.003). Full results are reported in Table S8 in Multimedia Appendix 1, and the analysis specifications are provided in Multimedia Appendix 2.

Robustness Checks

Examination-day fixed effects (+2.79 points per SD, 95% CI 2.23-3.34), case-by-day cell fixed effects (+2.78 points per SD, 95% CI 2.23-3.33), and recorded station-slot-label fixed effects (+2.83 points per SD, 95% CI 2.25-3.40) produced estimates similar to those of the primary analysis. Similar results were obtained in grouped cross-validation, leaving out 1 class, 1 examination day, or 1 recorded station-slot label at a time. These checks assessed predictive stability across administrative conditions but did not estimate examiner-specific effects or establish homogeneity across days (see Table S1 in Multimedia Appendix 1; also see Multimedia Appendix 2).

Discussion

Principal Findings

This prospectively sequenced observational study examined whether preexamination AMTES performance was associated with the recorded score on an independently assessed OSCE history-taking station, whether the association depended on including the complaint-matched training-case score, and whether a model incorporating the platform score retained predictive performance in held-out students. Building on prior AMTES research on system development, scoring performance, learner acceptance, and educational effectiveness, this study extends the evaluation of routinely generated AMTES scores to their relationship with an independently administered OSCE under required curricular use. Platform performance was positively associated with the score recorded on a single, independently scored OSCE history-taking station after accounting for differences among the assigned OSCE cases. The association was also observed in the AMTES-unmatched presenting-complaint stratum and when the contribution of the presenting-complaint-matched training case was omitted from the platform score. The association attenuated but persisted when the predictor was reconstructed using only practices completed before the first examination day. Models incorporating the platform score retained predictive performance in internal cross-validation and added information beyond the 2 prior physical-examination course scores. The association also persisted after adjustment for those scores, practice volume, and case breadth. Together, these findings provide cohort-scale criterion-related evidence to inform the interpretation of AMTES scores in this curricular context.

Interpretation and Scope of the Evidence

The principal findings support a specific and bounded interpretation of AMTES scores under the current curricular deployment. They provide initial criterion-related evidence for interpreting routinely generated AMTES scores as a candidate indicator of between-student differences relevant to performance on a single, independently scored local OSCE history-taking station; they do not validate AMTES as a general measure of history-taking competence [10,31]. Score interpretations and uses require evidence specific to the proposed context and purpose [10,11,13,31]. The stronger observed association with history content than with the communication block on a common standardized scale is consistent with the content focus of the AMTES rubric. However, standardization addresses differences in scale but does not account for measurement error or the ceiling-prone distribution of communication scores.

The associations observed in the presenting-complaint stratum unmatched to AMTES and with the leave-matched-case-out score suggest that the overall association did not depend on including the complaint-matched AMTES case score. However, the leave-matched-case-out score remained highly correlated with the overall mean platform score, and these analyses compared different students rather than changes within the same students. Therefore, they do not establish within-student transfer of history-taking performance across clinical presentations.

After adjustment for both prior physical-examination course scores, the platform score remained positively associated with the recorded OSCE score, although the association was attenuated. This attenuation suggests that AMTES performance overlaps partly with measured prior physical-examination performance, while the remaining association suggests that the platform score provides information not captured by these prior assessments. However, because the earlier examinations are imperfect proxies for baseline history-taking ability, the residual association cannot be attributed specifically to learning through AMTES or fully separated from unmeasured learner characteristics. Accordingly, these findings are best interpreted as evidence of incremental predictive information within this cohort rather than as evidence of an independent educational effect of AMTES. Whether this additional information improves individual-level educational decisions remains untested. The absence of further average predictive gain from adding practice volume and case breadth should also be interpreted cautiously because both measures showed limited variation under the required 6-case assignment. This finding therefore does not establish that additional or broader practice lacks educational value.

The association persisted when the predictor was restricted to records submitted before the examination period, although it attenuated modestly. This attenuation could reflect fewer observations, the exclusion of later learning or examination-informed practice, or a combination of these mechanisms; the study design cannot distinguish among these explanations. Moreover, persistence of the association under the earlier cut-off does not exclude the possibility that students subsequently received examination information that affected their OSCE performance.

The association between platform and OSCE scores was observed in both sexes, with no detectable difference in slope. However, the interaction estimate was imprecise and does not establish equivalent performance. At the same platform score

XSL•FO
RenderX

and assigned OSCE case, women had a higher conditional mean OSCE score. The model without assigned OSCE-case information showed a smaller sex-related difference in mean out-of-fold prediction error; these results reflect different comparisons. The findings are therefore preliminary and do not establish fairness. Evaluation of subgroup performance, calibration, and the consequences of any score-based decisions will be needed before operational use [32,33].

Comparison With Prior Work

Prior studies of LLM-based virtual patients have primarily examined feasibility, simulated-patient fidelity, learner experience, and agreement between automated and human ratings [1,2]. Previous AMTES studies addressed platform scoring stability and human-AI consistency [3], educational effectiveness under voluntary extracurricular use [26], and learner acceptance [27]. Educational effectiveness and criterion-related evidence address complementary but distinct questions: the former concerns whether an intervention improves outcomes, whereas the latter concerns how the scores generated by the platform should be interpreted.

Intervention studies provide relevant neighboring evidence. A nonrandomized study reported higher preclerkship OSCE interview scores among students who participated in AI-simulated-patient training than among historical controls [25]. A recent quasi-experimental study similarly reported higher OSCE-based interview scores after AI-generated feedback following social-robotic virtual-patient interactions [34]. Both studies evaluated the effects of training or feedback interventions; neither examined whether automated scores accumulated during routine practice corresponded to performance on an independently scored clinical examination. A naturalistic study documented intensive voluntary use of an AI-based OSCE platform during national licensure examination preparation but did not link its routine automated scores to an independent examination outcome. The authors called for controlled or externally benchmarked studies of clinical competence and examination outcomes [35].

More broadly, the technology-enhanced simulation literature has synthesized validity evidence for simulation-based assessment [36], and a study in a conventional problem-based learning curriculum found positive correlations between formative-assessment scores and OSCE and progress-test scores [37]. However, this literature does not establish whether scores generated during cohort-scale curricular use of an LLM-based history-taking trainer correspond to performance on an independently scored OSCE station. This study helps address this gap by linking routinely generated AMTES scores to an external, assessor-blinded OSCE criterion under required curricular use.

Implications for Education and Future Research

The criterion-related and internally cross-validated evidence supports further evaluation of AI-trainer analytics as a candidate scalable, low-stakes formative signal [19,20]. Such analytics could complement, but should not replace, faculty-observed performance assessments within a program of assessment [38].

One potential human-in-the-loop workflow, not tested in this study, would use persistently low platform scores available before an individual student's examination as a low-stakes triage signal for faculty review, rather than as a basis for competence classification or other consequential assessment decisions. Faculty could examine item-level feedback and patterns across cases, confirm any concerns through direct observation or discussion with the student, and offer targeted support, such as focused questioning practice, supervised simulation, or additional feedback. The platform signal would inform faculty judgment; it would not diagnose competence, automatically trigger a consequential decision, or replace direct assessment. Before implementation, evidence would be needed that flagged learners benefit from the support offered and that acting on the signal improves subsequent performance. This study addresses neither question; moreover, the remaining prediction error and the absence of an evaluated decision rule limit immediate individual-level use.

These findings motivate 3 priorities for future research. First, the association between platform scores and the criterion should be prospectively evaluated in new cohorts, institutions, platform versions, and case sets. Second, candidate formative workflows should be evaluated to determine whether faculty review and targeted support initiated by the platform signal improve subsequent learning or clinical performance. Such evaluations should also examine differential performance, subgroup calibration, and potential bias across learner groups beyond the single characteristic examined here. Third, randomized studies of trainer exposure or practice dose are needed to determine whether platform practice itself improves independently assessed history-taking performance. Any alert or decision rule would require separate evidence supporting the specific interpretation and use of the scores [10,11].

Strengths

A key design strength was the course-required deployment within an eligibility-defined curricular cohort. This approach reduced dependence on voluntary platform adoption while preserving variation in preexamination practice behavior. Near-complete record linkage further supported cohort coverage.

The student-specific time stamp restriction and the separately administered OSCE were additional design features. The criterion was independent of the platform in 2 scenarios: it was obtained using a separate assessment instrument whose local rubric was referenced to China's National Medical Licensing Examination, and it was scored by examiners who had no access to students' platform records or scores. This reduces concern that the observed association arose solely from the platform's scoring process or from examiners being influenced by prior knowledge of platform performance. The association was also evaluated under routine curricular examination conditions involving multiple cases, classes, and examination days, rather than under research-optimized assessment conditions.

Under the availability-driven allocation process, examination staff assigned students to available stations, and students did not choose whether they entered the AMTES-unmatched presenting-complaint stratum. Although this process did not constitute formal randomization, it reduced the specific concern

XSL•FO
RenderX

that students selected their OSCE case according to their own preparation or preferences.

Limitations

Overview

This study has several limitations, and these are described in the following subsections.

Observational Design and Residual Confounding

As this was a prospectively sequenced observational study rather than a randomized intervention, residual confounding cannot be fully ruled out. Measured prior physical-examination performance was an imperfect proxy for baseline history-taking ability, and unmeasured differences in learner ability, motivation, or engagement may have contributed to both AMTES and OSCE performance.

OSCE Measurement and Administration

Each student completed only 1 OSCE station, and each performance received a single examiner rating. Consequently, interrater reliability and student-level cross-station generalizability could not be estimated in this study [9,30]. OSCE checklists may also incompletely represent increasing levels of clinical expertise [39]. Case fixed effects accounted for average score differences among the assigned cases but could not separate case, examiner, standardized-patient, station, and session effects. Although examiners were not given access to platform records, random rating error, systematic scoring differences, and possible effects of examiners' prior familiarity with students could not be excluded. Examiner-specific effects could not be estimated because examiner identity and formal cross-examiner calibration data were unavailable. Random rating error would reduce outcome reliability, potentially attenuating score correlations and reducing statistical precision [9], whereas systematic examiner-, case-, or session-related effects could bias the association in either direction. Disattenuated coefficients were therefore not estimated because the required reliability parameters were unavailable.

The recorded station-slot labels represented composite administrative conditions rather than verified examiner identifiers and therefore could not isolate examiner-specific effects. Sensitivity analyses using these labels, examination day, class, and case-by-day cells produced similar associations, but they cannot replace repeated ratings or a crossed station-by-rater design. As the OSCE was administered over 5 days, students examined later had more time to practice, and unrecorded exchange of examination information across days could not be excluded. With only 5 examination days and 9 classes, cluster-level inference was not used for the primary CIs because estimates based on so few clusters would have been unstable. The HC3 CIs therefore relied on student-level independence, whereas grouped cross-validation assessed predictive stability rather than correcting statistical inference for clustering.

Incomplete Record Linkage

A small number of eligible OSCE examinees could not be linked to platform records, and 1 linked examinee had no analyzable preexamination practice. The comparison of OSCE scores between linked and unlinked examinees was imprecise because the unlinked group was small. Residual linkage bias therefore cannot be excluded.

Limited Generalizability and Incomplete Measurement Evidence

The study was conducted at a single institution using 1 cohort, 1 platform version, and 1 OSCE case set. Internal cross-validation used held-out students from the same cohort and therefore does not establish generalizability across institutions, cohorts, platform versions, training cases, OSCE cases, raters, stations, or assessment occasions. Temporal external validation or multisite validation is required before predictive use. Predictor-side item-level scoring properties and item-level construct alignment were not reevaluated, and reproducibility across platform versions and case sets remains unknown. As the trainer's scoring behavior depends on the configured models, prompt templates, case content, and scoring rubrics, replication studies must specify these elements [3]. Sex was the only demographic characteristic available for subgroup analysis. Although no difference in the platform-score slope by sex was detected, the CI was wide and does not establish equivalent performance. Age was not recorded in the source data, and no other demographic or subgroup characteristic was available. Model performance across other learner groups, as well as the consequences of any proposed alert threshold, therefore requires evaluation before operational use.

Operational Decision Use Was Not Evaluated

The study did not develop or prospectively evaluate an alert, classification, or decision rule. No threshold was selected, and sensitivity, specificity, predictive values, alert burden, and the consequences of false-positive and false-negative classifications were not assessed. Individual-level prediction precision was also limited (Table S5 in Multimedia Appendix 1). The findings therefore do not support automated alerts, summative assessment, progression decisions, or other high-stakes uses.

Developer-Evaluator Dual Role

Several authors developed or deployed AMTES (see the "Conflicts of Interest" section), and the system was evaluated at their own institution and within their own curriculum, creating the potential for setting-specific optimization and favorable interpretation. Design features that limit, but do not eliminate, this risk are described in the "Conflicts of Interest" section. Independent external replication by investigators not involved in developing or deploying the platform is required before broader claims can be made.

Conclusions

This study provides cohort-scale criterion-related evidence linking routine AMTES practice scores to recorded scores on an independently scored local OSCE history-taking station under required curricular use. The association did not depend on including the complaint-matched training-case score and attenuated but persisted under a cohort-wide preexamination cut-off. AMTES scores also added predictive information beyond the 2 prior physical-examination course scores in internal cross-validation. These between-student findings do

XSL•FO
RenderX

not establish within-student transfer or causal effects of platform practice, and residual confounding by prior ability may remain.

These findings support further evaluation of AMTES analytics as a candidate low-stakes formative signal for faculty review and targeted support. Independent external validation and prospective evaluation of the proposed workflow are needed before operational use. Any faculty-facing signal should trigger further assessment and support rather than classification of competence, and should complement direct performance assessment.

Funding

We acknowledge support from the Key Project on Medical Education Research by the National Medical Education Development Center, Medical Education Branch of the Chinese Medical Association (grant 2025A25), and the 2025 Teaching Reform and Research Project (AI Curriculum Special Program) of Shantou University Medical College (Number 8). The funders had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

Deidentified derived analytic data underlying all reported tables and figures, along with the Python analysis and figure-generation scripts, are available from the corresponding author (CS) on reasonable request. As raw platform interaction logs from this single-institution cohort may contain contextual details that increase the risk of reidentification, those materials are not publicly released. A redacted version may be made available on reasonable request, subject to institutional and ethics requirements.

Generative AI Use

The authors declare that generative AI was used in the creation of this manuscript. Generative AI (ChatGPT, Claude) was used for proofreading, editing, or language polishing; translation. The authors verified all outputs of the tool(s) used in the preparation of this manuscript and take full responsibility for the final content of the manuscript.

Authors' Contributions

CS developed the methodology, secured funding, supervised the research team, and led project management. OX conceptualized and designed the study and led project management. WZ conceptualized and designed the study, developed the software system, contributed to data analysis, and prepared the original manuscript draft. YL developed the research methodology, conducted the investigation, performed data analysis, and prepared the original manuscript draft. YZ developed the analytical methodology, performed formal statistical analysis, and generated data visualizations. LW contributed to the conceptualization and design of the study and assisted with project management. XH contributed to the investigation and data curation. XL contributed to the conceptual framework and study design. MY was responsible for data curation and contributed to data analysis. WZ and YL contributed equally to this work and are cofirst authors; CS and OX contributed equally to this work and are cocorresponding authors. All authors reviewed and approved the final manuscript for publication.

Conflicts of Interest

No financial competing interests are declared. Several authors were involved in the development or institutional deployment of AMTES, the platform evaluated in this study, which represents a relevant nonfinancial competing interest (developer/evaluator dual role). The individual roles were as follows: WZ developed the AMTES software system. WZ and YZ were responsible for its institutional deployment and day-to-day operation during the study period; neither took part in OSCE administration or scoring. YL and LW served as OSCE examiners in 3 and 2 of the 9 examination sessions, respectively; like all examiners, they did not know in advance which students they would examine and had no access to individual AMTES practice records or scores when scoring. YZ performed the formal statistical analysis of the linked dataset. The remaining authors (XL, MY, XH, OX, and CS) had no role in AMTES development, deployment, or operation, or in the OSCE. Design features that limit, but do not eliminate, the influence of this dual role are described in the “Methods” section: the OSCE used a separate instrument and was administered and scored outside AMTES by examiners with no access to platform records or scores; the outcome was the officially recorded station total; practice scoring was fully automated; and the linked data were pseudonymized before analysis. These features do not remove the need for independent external replication.

Multimedia Appendix 1

Additional data analysis.
[DOCX File , 326 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Detailed study methods and results.
[DOCX File , 33 KB-Multimedia Appendix 2]

XSL•FO
RenderX

References

1. Holderried F, Stegemann–Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A Generative Pretrained Transformer (GPT)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ. Jan 16, 2024;10:e53961. [doi: 10.2196/53961] [Medline: 38227363]
2. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model–powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ. Aug 16, 2024;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
3. Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, et al. Development and validation of a large language model–based system for medical history-taking training: prospective multicase study on evaluation stability, human-AI consistency, and transparency. JMIR Med Educ. Aug 29, 2025;11:e73419. [doi: 10.2196/73419] [Medline: 40882613]
4. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Medical Education. Mar 18, 2009;43(4):303-311. [doi: 10.1111/j.1365-2923.2008.03286.x] [Medline: 19335571]
5. 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. Jul 02, 2019;21(7):e14676. [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
6. 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. Sep 07, 2011;306(9):978-988. [doi: 10.1001/jama.2011.1234] [Medline: 21900138]
7. Zeng J, Qi W, Shen S, Liu X, Li S, Wang B, et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. J Med Internet Res. Nov 13, 2025;27:e79091. [FREE Full text] [doi: 10.2196/79091] [Medline: 41232097]
8. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. Sep 2003;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x] [Medline: 14506816]
9. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. Sep 2004;38(9):1006-1012. [doi: 10.1111/j.1365-2929.2004.01932.x] [Medline: 15327684]
10. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. Sep 1995;50(9):741-749. [doi: 10.1037/0003-066x.50.9.741]
11. Kane MT. Validating the interpretations and uses of test scores. J Educational Measurement. Mar 14, 2013;50(1):1-73. [doi: 10.1111/jedm.12000]
12. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. Feb 2006;119(2):166.e7-166.e16. [doi: 10.1016/j.amjmed.2005.10.036] [Medline: 16443422]
13. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. Jun 2015;49(6):560-575. [doi: 10.1111/medu.12678] [Medline: 25989405]
14. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. Jan 1979;13(1):41-54. [Medline: 763183]
15. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. Mar 22, 1975;1(5955):447-451. [FREE Full text] [doi: 10.1136/bmj.1.5955.447] [Medline: 1115966]
16. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach. Sep 2013;35(9):e1437-e1446. [doi: 10.3109/0142159X.2013.818634] [Medline: 23968323]
17. Pell G, Fuller R, Homer M, Roberts T, International Association for Medical Education. How to measure the quality of the OSCE: a review of metrics - AMEE guide no. 49. Med Teach. 2010;32(10):802-811. [FREE Full text] [doi: 10.3109/0142159X.2010.507716] [Medline: 20854155]
18. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. Mar 2004;38(2):199-203. [doi: 10.1111/j.1365-2923.2004.01755.x] [Medline: 14871390]
19. Ellaway RH, Pusic MV, Galbraith RM, Cameron T. Developing the role of big data and analytics in health professional education. Med Teach. Mar 2014;36(3):216-222. [doi: 10.3109/0142159X.2014.874553] [Medline: 24491226]
20. Chan T, Sebok-Syer S, Thoma B, Wise A, Sherbino J, Pusic M. Learning analytics in medical education assessment: the past, the present, and the future. AEM Educ Train. Apr 2018;2(2):178-187. [FREE Full text] [doi: 10.1002/aet2.10087] [Medline: 30051086]
21. Gierl MJ, Latifi S, Lai H, Boulais AP, De Champlain A. Automated essay scoring and the future of educational assessment in medical education. Med Educ. Oct 2014;48(10):950-962. [doi: 10.1111/medu.12517] [Medline: 25200016]
22. Salt J, Harik P, Barone MA. Leveraging natural language processing: toward computer-assisted scoring of patient notes in the USMLE step 2 clinical skills exam. Acad Med. Mar 2019;94(3):314-316. [doi: 10.1097/ACM.0000000000002558] [Medline: 30540567]
23. Cross J, Kayalackakom T, Robinson RE, Vaughans A, Sebastian R, Hood R, et al. Assessing ChatGPT's capability as a new age standardized patient: qualitative study. JMIR Med Educ. May 20, 2025;11:e63353. [FREE Full text] [doi: 10.2196/63353] [Medline: 40393017]

XSL•FO
RenderX

24. Lee HY, Kim J, Choi H, Bae H, Jeong A, Choi S, et al. Comparing AI chatbot simulation and peer role-play for OSCE preparation: a pilot randomized controlled trial. BMC Med Educ. Nov 24, 2025;25(1):1755. [FREE Full text] [doi: 10.1186/s12909-025-08308-y] [Medline: 41286823]
25. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ. Sep 23, 2024;10:e58753. [FREE Full text] [doi: 10.2196/58753] [Medline: 39312284]
26. Liu Y, Zhu Y, Zhang W, Lu X, Wu L, Yue M, et al. Real-world impact and educational effectiveness of an AI-powered medical history-taking system: retrospective propensity score-matched cohort study. JMIR Med Educ. Mar 24, 2026;12:e89367. [FREE Full text] [doi: 10.2196/89367] [Medline: 41734000]
27. Liu Y, Zhu Y, Shi C, Lu X, Wu L, Yue M, et al. User acceptance of an AI-powered medical history-taking training system among undergraduate medical students: mixed methods study. JMIR Med Educ. Jul 21, 2026;12:e101425. [FREE Full text] [doi: 10.2196/101425] [Medline: 42387330]
28. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. Oct 20, 2007;370(9596):1453-1457. [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
29. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. Adv Health Sci Educ Theory Pract. 2003;8(1):27-47. [doi: 10.1023/a:1022639521218] [Medline: 12652167]
30. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. Med Educ. Oct 2002;36(10):972-978. [doi: 10.1046/j.1365-2923.2002.01320.x] [Medline: 12390466]
31. St-Onge C, Young M, Eva KW, Hodges B. Validity: one word with a plurality of meanings. Adv Health Sci Educ Theory Pract. Oct 2017;22(4):853-867. [doi: 10.1007/s10459-016-9716-3] [Medline: 27696103]
32. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC. American Educational Research Association; 2014.
33. Zwick R. Fairness in educational measurement: theory and concepts. In: Cook LL, Pitoniak MJ, editors. Educational Measurement. 5th ed. New York, NY. Oxford University Press; 2025:383-438.
34. Borg A, Schiött J, Ivegren W, Gentline C, Huss V, Hugelius AM, et al. AI-generated feedback following social robotic virtual patient interactions and medical student performance: nonrandomized quasi-experimental study. JMIR Med Educ. Mar 25, 2026;12:e90368. [FREE Full text] [doi: 10.2196/90368] [Medline: 41881044]
35. Song JW, Kim M, Lee YG, Cheong SW, Kim JJ, Cho J, et al. Naturalistic adoption and deliberate practice use of an AI-based OSCE platform during national licensure preparation. NPJ Digit Med. Jul 15, 2026. [FREE Full text] [doi: 10.1038/s41746-026-03024-3] [Medline: 42458025]
36. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med. Jun 2013;88(6):872-883. [doi: 10.1097/ACM.0b013e31828ffdcf] [Medline: 23619073]
37. Couto LB, Durand MT, Wolff ACD, Restini CBA, Faria M, Romão GS, et al. Formative assessment scores in tutorial sessions correlates with OSCE and progress testing scores in a PBL medical curriculum. Med Educ Online. Dec 2019;24(1):1560862. [FREE Full text] [doi: 10.1080/10872981.2018.1560862] [Medline: 31023185]
38. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract. Jan 1996;1(1):41-67. [doi: 10.1007/BF00596229] [Medline: 24178994]
39. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. Oct 1999;74(10):1129-1134. [doi: 10.1097/00001888-199910000-00017] [Medline: 10536636]

Abbreviations

AMTES: AI-Powered Medical History-Taking Training and Evaluation System
GAIDeT: Generative AI Delegation Taxonomy
GenAI: generative AI
HC3: heteroskedasticity-consistent SEs (type 3)
LLM: large language model
MAE: mean absolute error
OSCE: objective structured clinical examination
RQ: research question
SP: standardized patient
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

XSL•FO
RenderX

Edited by T de Azevedo Cardoso; submitted 16.Aug.2026; peer-reviewed by B Depboylu, Y Ohira, A Hudon; comments to author 09.Sep.2026; revised version received 22.Sep.2026; accepted 20.Oct.2026; published 03.Nov.2026

Please cite as:
Zhang W, Liu Y, Zhu Y, Lu X, Wu L, Yue M, Hong X, Xia O, Shi C
Association of Preexamination AI Virtual-Patient Trainer Scores With Scores on a Local Single-Station History-Taking Objective Structured Clinical Examination: Observational Cohort Study
JMIR Med Educ 2026;12:e109752
URL: https://mededu.jmir.org/2026/1/e109752
doi: 10.2196/109752
PMID:

©Weishan Zhang, Yang Liu, Yiying Zhu, Xian Lu, Liping Wu, Minghui Yue, Xiaolin Hong, Oudong Xia, Chujun Shi. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX