Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Effect of Competency-Oriented Multimodal Training on Standardized Patient Performance: A Prospective, Assessor-Blinded, Randomized Controlled Mixed-Methods Study

MingYue Zhou
the Seventh Affiliated Hospital of Sun Yat-sen University
Yawen Zhang
the Seventh Affiliated Hospital of Sun Yat-sen University
Qian Chen
the Seventh Affiliated Hospital of Sun Yat-sen University
Jingfang Luo
the Seventh Affiliated Hospital of Sun Yat-sen University
Xin Shu
the Seventh Affiliated Hospital of Sun Yat-sen University
Zhenning Liang
the Seventh Affiliated Hospital of Sun Yat-sen University
Mingqing Zhao
the Seventh Affiliated Hospital of Sun Yat-sen University
Jiancheng Wang
wangjch38@mail.sysu.edu.cn

the Seventh Affiliated Hospital of Sun Yat-sen University

Research Article

Keywords: Standardized patient, Simulation-based education, Training evaluation, Medical education, Mixed-methods research

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11394618/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Effect of Competency-Oriented Multimodal Training on Standardized Patient Performance: A Prospective, Assessor-Blinded, Randomized Controlled Mixed-Methods Study

MingYue Zhou[1, #], Yawen Zhang[1, #], Qian Chen[1], Jingfang Luo[1], Xin Shu[1], Zhenning Liang[1], Mingqing Zhao[1], Jiancheng Wang[1, *]

Affiliations

[1] Scientific Research Center, the Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen 518107, P. R. China

[#]**These authors contributed equally**

[*]**Corresponding Author:**

E-mail: wangjch38@mail.sysu.edu.cn (Prof. Wang)

Abstract

Background

Standardized patients (SPs) support clinical teaching and assessment, but their value depends on authentic, accurate, and consistent performance and effective feedback. This mixed-methods research plan compares competency-oriented multimodal training (Combi-SP) with conventional training and explores SPs' experiences of the training.

Methods

This study recruited 64 SPs, equally allocated to either Combi-SP or conventional training, with eight 45-minute class hours for each group. Objective performance was assessed at baseline, immediately post-training, and at six weeks via blinded video ratings. Linear mixed-effects models were used to evaluate changes in performance. A qualitative component was incorporated into the immediate post-training assessment: two open-ended questionnaire items investigated helpful training components, persistent challenges, and support requirements. Responses from 59 out of 63 questionnaire respondents were analysed using qualitative content analysis and synthesised with quantitative findings in a joint display.

Results

In this study, 63 SPs completed the immediate assessment and 60 completed follow-up. Combi-SP showed greater adjusted improvement in objective performance at both time points: 4.71 points immediately after training (95% CI, 2.25–7.16) and 4.43 points at six weeks (95% CI, 1.93–6.93). The four performance domains and self-efficacy also favored Combi-SP. Overall interactions for the two error-count outcomes were not statistically significant. Three qualitative themes concerned understanding the case and controlling disclosure, specific feedback and targeted practice, and training demands and refresher support. The joint display linked the score patterns to these experiences while retaining uncertainty about error reduction and longer-term transfer.

Conclusions

Competency-oriented multimodal training produced greater improvements in objective SP performance than conventional training, with part of the advantage retained at 6 weeks.

Keywords

Standardized patient, Simulation-based education, Training evaluation, Medical education, Mixed-methods research

Background

Standardized patients (SPs) are individuals who are systematically trained to portray a specific patient history, symptoms, emotions, and behavioral responses in a reproducible and controlled manner for teaching or assessment.[1-3] Without exposing real patients to additional risk, SPs can support training in history taking, communication, clinical reasoning, professionalism, and physical examination.[2] Recent applications include early training in psychiatric interviewing and communication.[4] SPs may also participate in educational assessment as observers, raters, and feedback providers.[1] In undergraduate clinical education, SPs help address shortages of representative cases, uneven access to clinical learning opportunities, and limited availability of real patients.[5] They are therefore widely used in clinical skills courses and objective structured clinical examinations (OSCEs).[6]

The educational value of SP-based teaching depends not only on whether SPs are used, but also on whether their performance is authentic, accurate, and consistent. Omitting key case information, volunteering information that should be elicited by the learner, changing symptom intensity across encounters, or rating performance on the basis of global impressions rather than observable behavior can introduce contextual error and compromise both the learning experience and assessment fairness.[7] The Association of Standardized Patient Educators (ASPE) Standards of Best Practice emphasize safe work environments, case development, role portrayal training,

performance and scoring quality, program management, and professional development. These principles are also reflected in minimum standards for SP programs.[1,8] Instruments such as the Maastricht Assessment of Simulated Patients (MaSP) identify role authenticity and feedback quality as important components of SP performance.[9,10] Transparent reporting of SP characteristics, training, encounters, and assessment procedures is necessary to judge and replicate SP research.[11]

Previous medical education research has often treated SPs as an instructional tool and compared students taught with SPs with those receiving traditional instruction in terms of knowledge, clinical skills, communication, or satisfaction. For example, a Chinese study reported better performance in history taking, physical examination, and clinical skills after SP-assisted clinical practice teaching, although theoretical test scores did not significantly differ. A broader meta-analysis of technology-enhanced simulation also reported benefits for learners' knowledge, skills, and behaviors compared with no intervention.[12] Multimodal approaches have additionally been evaluated in pathophysiology teaching, with improvements in selected domains of student performance.[13] Such learner-focused studies support educational innovation but do not answer a different question: what type of training improves the competencies of SPs themselves? When a research team cannot or should not alter undergraduate course grades, making SPs the study population and standardized SP performance the primary outcome provides a more direct causal link between the intervention and the outcome.

Existing evidence suggests that structured training, video review, self-assessment, peer and faculty feedback, and calibration across multiple SPs can support SP performance and rating consistency.[14-16] Multicenter studies of SP standardization further indicate that authenticity alone is not sufficient; the information provided by SPs should also be checked against the case script.[16] Students can evaluate SP performance from the learner's perspective after each encounter, and a student version of an SP performance rating scale has been developed and psychometrically

evaluated.[17] Training student SPs for OSCEs has also been shown to be feasible, although this population differs from established institutional SP pools[18]. Survey evidence shows substantial variation in SP program organization and training practices[19]. An important remaining question is whether integrated, competency-oriented training is superior to conventional SP training under blinded assessment and whether any performance advantage persists after training.

Competency-based medical education emphasizes explicit outcomes and the development of observable capabilities[20]. Broader simulation reviews identify feedback, repeated practice, and multiple learning strategies as useful features of educational design[21,22]. Deliberate practice focuses on targeted repetition with feedback to address performance deficits[23,24]. These findings provide a rationale for combining training strategies, while the effectiveness of such a combination for SPs requires direct evaluation.

In this study, we developed a competency-oriented Combi-SP training program that integrates structured case interpretation, symptom and emotion portrayal, management of unscripted questions, video-assisted deliberate practice, peer calibration, and assessment and feedback training into a continuous training pathway. The program was compared with conventional SP training of equal duration. The primary objective was to compare changes in objective SP performance immediately after training and at six-week follow-up. Secondary objectives were to examine case-information errors, role-maintenance deviations, inter-encounter consistency, feedback quality, and self-efficacy. Undergraduate course grades were not study outcomes; students served only as anonymous experiential raters so that routine teaching assessment and student grades were not affected.

Methods

Study design and setting

This study was a single-center, prospective, parallel-group, assessor-blinded, wait-list-controlled randomized mixed-methods trial. It was jointly conducted by the Department of Pathophysiology, School of Medicine, Sun Yat-sen University, the Office of Academic Affairs of the Seventh Affiliated Hospital of Sun Yat-sen University, and the Clinical Skills Center, with a duration of 3 months. The quantitative component included three assessment points: before training (T0), immediately after training (T1), and 6 weeks after training (T2). The study flow is shown in **Fig. 1**.

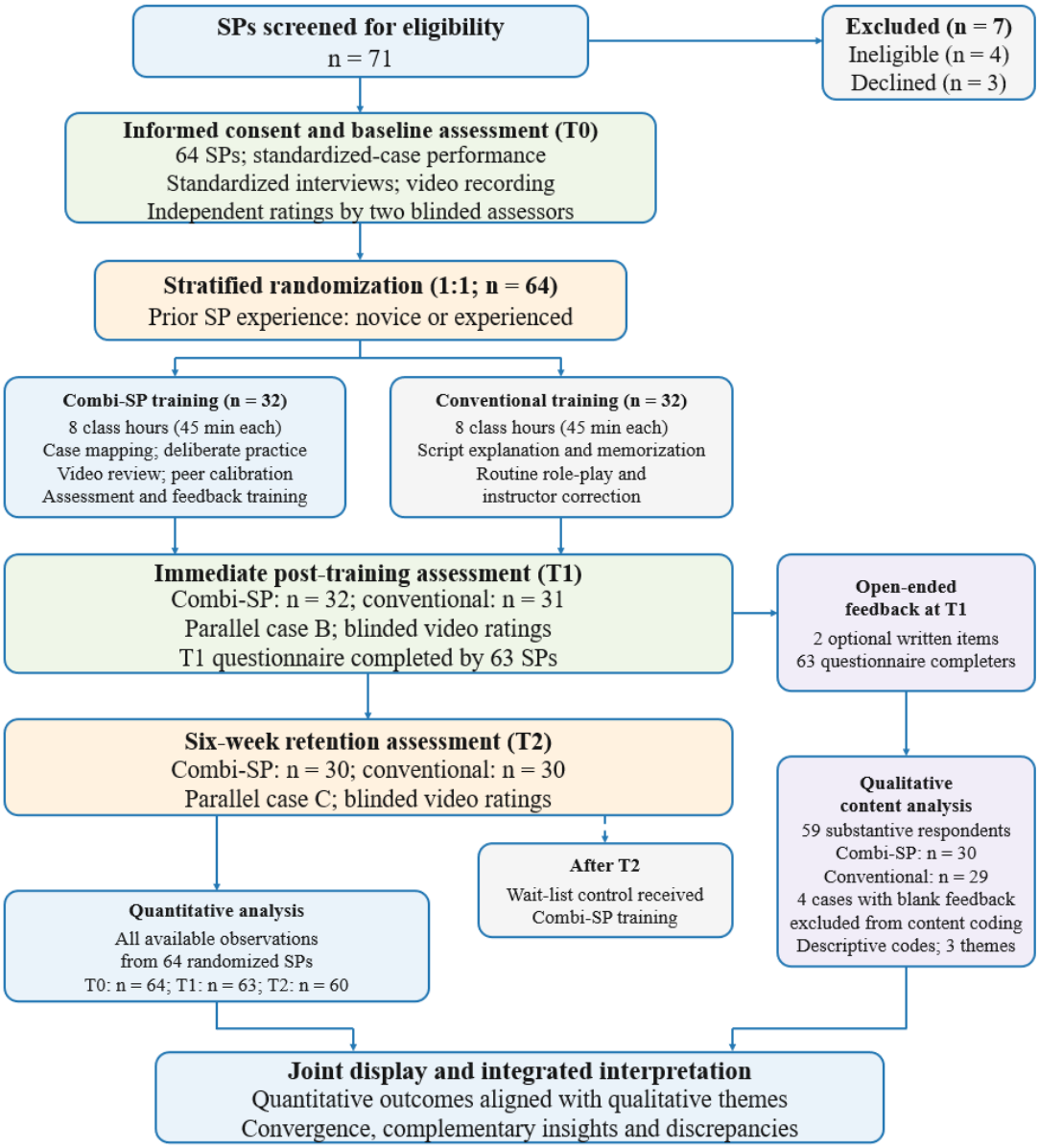


Fig.1 Study flow. SPs were assessed at baseline (T0), randomized after stratification by prior SP experience, and reassessed immediately after training (T1) and at six weeks (T2). Open-ended feedback collected at T1 underwent qualitative content analysis and was integrated with quantitative findings in a joint display.

Participants

Inclusion criteria

1. Age 18 years or older.
2. Already enrolled in the institution's SP pool or having passed the initial screening for the present SP recruitment.
3. Able to understand Chinese-language case scripts and complete the required role-play and feedback tasks.
4. Willing to participate in all training sessions, three video-recorded assessments, and questionnaire surveys.
5. Voluntary participation with signed informed consent.

Exclusion criteria

1. Physical or psychological reasons that made the assigned case unsuitable for the participant.
2. Refusal to be video-recorded or refusal to allow faculty to review the recordings for scoring.
3. Inability to complete the baseline assessment.
4. Participation during the study period in external advanced SP training that substantially overlaps with the study intervention.
5. Any other circumstance judged by the investigators to be likely to materially affect adherence to the intervention or outcome assessment.

Withdrawal and dropout criteria

Participants could withdraw unconditionally at any time. Failure to complete the primary outcome assessment at T1 was defined as early dropout; participants who completed T1 but not T2 were considered lost to follow-up. Withdrawal did not affect subsequent participation in routine SP teaching. The research team recorded the timing and reason for withdrawal when available, but participants were not required to provide a reason.

Sample size estimation

The primary effect was defined as the standardized mean difference in change in objective SP performance between the two groups. Based on the moderate-to-large effects reported in previous structured SP training studies, and considering that this was the first single-center validation of the program, an effect size of Cohen's $d = 0.75$ was prespecified, with a two-sided $\alpha = 0.05$ and power $(1-\beta) = 0.80$. An independent-group comparison required 29 participants per group. Allowing for approximately 10% attrition, the sample size was 64 participants, with 32 per group.

Randomization and allocation concealment

A researcher who was not involved in recruitment, training, or outcome rating generated the randomization sequence by computer. Randomization was stratified by prior SP experience: novice SPs (≤7 previous encounters) and experienced SPs (>7 encounters). Within each stratum, variable block randomization with block sizes of 4 or 6 allocated participants 1:1 to the Combi-SP or conventional training group. Assignments were placed in sequentially numbered, opaque, sealed envelopes and opened only after completion of the T0 assessment. To reduce contamination, the two groups were trained at different times or locations, and training materials were not shared with the other group before T2.

Blinding

Because the training formats differed, SPs and instructors could not be blinded. The two faculty members from the Office of Academic Affairs/Clinical Skills Center who rated the video recordings did not participate in training and did not have access to group assignments. Video files were recoded and randomly ordered, with information on training group and assessment time removed. The statistician had access only to Group A/Group B labels until the primary analysis was completed. If the difference between the two raters' total scores exceeded a prespecified threshold, 10% of the maximum total score, a third senior faculty member reviewed the recording.

Development of cases and assessment materials

Three parallel cases were jointly developed by clinical faculty, SP trainers, and medical education assessment specialists for T0, T1, and T2. Cases were selected from common internal medicine or general practice scenarios in undergraduate clinical training that required limited physical portrayal and could be presented repeatedly, such as chest pain, abdominal pain, dizziness, or chronic disease follow-up. The three cases were comparable in the number of information points, emotional complexity, information that must be elicited by questioning, information that may be volunteered, and the number of standardized unscripted questions.

Each case package included: (1) learner task instructions; (2) the SP's basic identity and patient story; (3) the chief complaint and timeline of the history of present illness; (4) past, personal, and family history; (5) positive and negative information; (6) information-trigger rules; (7) symptom and emotion intensity; (8) standardized nonverbal behavior; (9) boundary responses to unscripted questions; and (10) the scoring checklist and training notes. Each case first underwent expert content-validity review and was then piloted by SPs and standardized interviewers who did not participate in the formal study.

Training interventions

Both groups received a total of eight class hours of training, with each class hour lasting 45 minutes. Training was generally completed in two sessions to control both training duration and instructor contact time. The same case topics were used in both groups, but the presentation of training materials and the practice strategies differed. Intervention components, delivery procedures, and comparator activities are described below and in **Table 1** to support replication[25].

Table 1 Comparison of SP training content between the two groups

Component	Combi-SP group	Conventional training group
Training goals	Role authenticity; case accuracy; consistency; assessment and feedback skills	Master the case script and complete basic role-play
Case learning	Patient story map, timeline, information-trigger rules, and a list of information that must not be volunteered	Item-by-item instructor explanation; individual memorization
Symptom and emotion portrayal	Graded demonstrations, nonverbal behavior training, and correction of overacting	General demonstration and real-time correction
Unscripted questions	Standardized question bank; training in boundary responses that avoid fabrication while maintaining role	Ad hoc instructor guidance when problems arise
Practice format	Video review, repeated rehearsal, and goal-directed deliberate practice	Routine role-play
Feedback format	Self-assessment + peer feedback + behavior-focused instructor feedback	General verbal instructor feedback
Consistency calibration	Multiple SPs portray the same case and calibrate performance across encounters	No dedicated calibration session
Assessment and feedback training	Behavioral checklists, patient-perspective feedback, and avoidance of global-impression bias	Brief introduction to the rating form or no dedicated training
Safety management	Role boundaries, fatigue reporting, psychological safety, and de-roling	Routine precautions

Combi-SP group

Module 1: Role standards and safety. The different roles of SPs in teaching, assessment, scoring, and feedback were clarified. The module also covered informed consent,

confidentiality, respect, physical and psychological safety, work boundaries, and de-roling after an encounter.

Module 2: Structured script interpretation. Cases were converted into patient story maps and clinical history timelines. Information was categorized as information that must be volunteered, information that should be provided only when elicited, information that must not be added, and information for which a standardized boundary response may be used.

Module 3: Authenticity and symptom portrayal. Training covered language, speech rate, pauses, facial expression, posture, and emotional responses. Graded anchors were used to control the intensity of pain, anxiety, anger, or avoidance and to prevent theatrical or mechanical performance.

Module 4: Management of unscripted questions. Instructors introduced clinically plausible questions that were not covered by the script. SPs were trained to maintain the patient perspective, avoid fabrication, avoid cueing the correct answer, and use pre-agreed boundary responses.

Module 5: Video review and deliberate practice. Each SP completed a full simulated interview and then reviewed the recording for self-assessment using the rating form. Peers and instructors provided specific feedback, after which the SP repeated the encounter with one or two targeted improvement goals.

Module 6: Observation, assessment, and feedback. SPs learned to distinguish observable behavior from subjective impressions and to use behavioral checklists correctly. Feedback followed a "specific behavior - patient experience - actionable suggestion" framework. Feedback was directed at observed performance and intended to guide improvement, consistent with operational definitions of feedback in clinical education[26].

Module 7: Multi-SP calibration and quality control. Several SPs portrayed the same case. Trainers compared key information, emotional intensity, response pace, and rating differences to establish common performance standards and a list of recurrent problems.

We have provided a framework diagram for your reference, see **Fig. 2.**

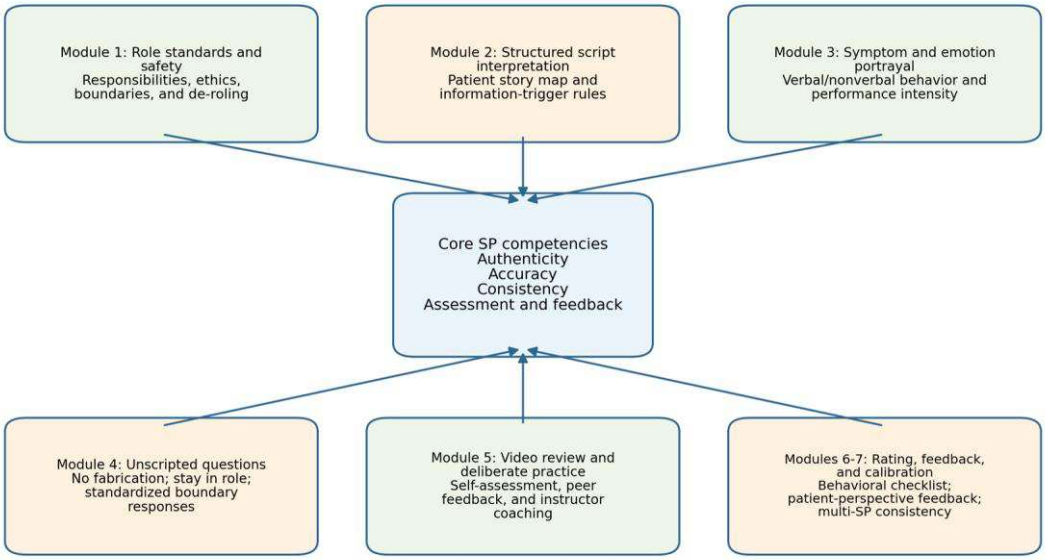


Fig.2 Combi-SP competency training framework. Seven modules targeted the core competencies of authenticity, accuracy, consistency, and assessment/feedback ability through structured script interpretation, portrayal training, management of unscripted questions, deliberate practice, and calibration.

Conventional training group

The conventional group received training of equal duration, including an introduction to basic SP concepts, item-by-item explanation of the case script, individual memorization, role rehearsal, and real-time instructor correction. This group did not use patient story maps, systematic training for unscripted questions, video review, peer calibration, or a behavior-focused feedback framework. After completion of T2, the conventional group received the complete Combi-SP program as a wait-list control.

Assessment procedures

Each assessment was conducted by the same pool of trained standardized interviewers using a fixed questioning framework, thereby minimizing confounding from differences in undergraduate interviewing ability. Each interview lasted 8–10

minutes, followed by a 3–5 minute SP rating or feedback task. All encounters were video-recorded in the same room using the same camera position, audio equipment, and time prompts. Standardized interviewers did not participate in SP training or video rating.

Undergraduate students interacted with SPs during teaching activities that did not contribute to course grades and completed a brief anonymous rating of SP authenticity. Student ratings were secondary outcomes only. They did not affect course grades and were not used for SP recruitment, remuneration, or dismissal. This was stated clearly during recruitment to reduce social desirability and evaluation pressure.

Outcome measures

Primary outcome

The primary outcome was the total objective SP performance score assigned by blinded faculty from the Office of Academic Affairs/Clinical Skills Center using the study-specific Objective SP Performance Rating Form. The initial form was developed from the ASPE Standards of Best Practice, the MaSP, previous SP training and standardization studies, and requirements for case-information consistency[1,9,14-17].It included four domains — role authenticity, case-information accuracy, consistency and professional behavior, and assessment and feedback skills — comprising 20 items rated from 1 to 5. Total scores ranged from 20 to 100, with higher scores indicating better SP performance.

Before use as the primary outcome instrument, 5 experts in medical education, clinical teaching, SP training, and educational measurement evaluated item relevance and clarity. Content validity evaluation considered the item-level content validity index (I-CVI) and the scale-level average content validity index (S-CVI/Ave)[27,28]. An additional 13 SPs who did not enter the formal trial completed pilot testing to assess comprehensibility, rating time, and preliminary inter-rater agreement. Internal

consistency was assessed using Cronbach's alpha, and inter-rater agreement was assessed using the intraclass correlation coefficient (ICC) [29,30].

Secondary outcomes

- Case-information accuracy: number of omitted key history items, instances of volunteering information without being asked, inconsistent responses, and fabrication of information outside the script. Two raters coded each event against the case reference standard.
- Role maintenance: number of clear breaks in role, instances of cueing the interviewer toward the answer, and marked deviations from the scripted intensity of symptoms or emotions.
- Consistency and professional behavior: the four-item domain of the Objective SP Performance Rating Form, with scores ranging from 4 to 20. No separate inter-encounter agreement rate is available in the supplied dataset.
- Assessment and feedback skills: the five-item domain of the Objective SP Performance Rating Form, covering specificity, patient perspective, correspondence with observable behavior, actionability, and respectfulness. Domain scores range from 5 to 25.
- SP self-efficacy: a 15-item investigator-developed questionnaire rated on a 1–5 scale, with total scores ranging from 15 to 75. Expert content-validity review and pilot testing were completed before formal use; the instrument was not described as an established validated scale.
- Training satisfaction and acceptability: six 1–5 ratings of content relevance, pace, practice opportunities, feedback quality, psychological safety, and willingness to continue. Their mean forms the satisfaction score. Two optional open-ended items ask about helpful or unhelpful elements and remaining difficulties or support needs.
- Physical and psychological burden on SPs: fatigue, discomfort, emotional burden, and difficulty de-roling were recorded.

- Anonymous undergraduate experience ratings: authenticity, coherence of information, naturalness of interaction, and helpfulness of feedback were rated without collecting or altering student grades.

Study outcomes and assessment time points are provided in **Table 2**.

Table 2 Study outcomes and assessment time points

Level	Outcome	Measurement	Time point	Primary analysis
Primary	Objective SP performance score	20 items, 20-100 points; blinded faculty video rating	T0, T1, T2	Group-by-time interaction
Secondary	Case-information errors	Counts of omissions, unsolicited disclosure, contradictions, and fabrication	T0, T1, T2	Negative Negative binomial GEE
Secondary	Role maintenance	Breaks in role, answer cueing, intensity deviations	T0, T1, T2	Negative binomial GEE
Secondary	Four performance domains	Authenticity, accuracy, consistency/professional behavior, and assessment/feedback domain scores	T0, T1, T2	Linear mixed-effects model
Secondary	SP self-efficacy	15 items, 15-75 points	T0, T1, T2	Linear mixed-effects model
Secondary	Training satisfaction	Mean of six 1–5 ratings	T1	Descriptive analysis
Secondary	Undergraduate experience rating	Mean of six 1–5 ratings	T1, T2	Gaussian GEE clustered by SP
Secondary	Training burden	Mean of four 1–5 ratings	T1	Descriptive analysis
Qualitative	Open-ended training feedback	Two optional written items; coded themes and quotations	T1	Qualitative content analysis
Integration	Joint display	Related quantitative findings aligned with qualitative themes	After analysis	Support, extension and limits

Development and quality control of questionnaires and rating instruments

Investigator-developed instruments included the baseline and training-needs questionnaire, SP self-efficacy questionnaire, Objective SP Performance Rating Form, undergraduate experience rating form, and training satisfaction and burden questionnaire with two open-ended items. Item pools are based on the literature and ASPE standards. Expert review and pilot testing will assess coverage, clarity, and feasibility before empirical use.

Before formal video rating, the two raters completed frame-of-reference training using 3–5 recordings that were excluded from the formal analysis, with discussion of item definitions and score anchors. During formal scoring, raters worked independently and did not view each other's scores. Approximately one duplicate recording was inserted after every 20 recordings to monitor rater drift.

Data collection and management

Each SP was assigned a unique study identifier. The identity linkage file was stored separately from research data, and video files were encrypted with strict access restrictions. Electronic questionnaires included required quantitative fields, optional open-ended fields, plausible-value checks, and single-submission controls. All open-ended feedback was labelled with the same study identifier but stored independently of the identity linkage file. All identifying details were removed prior to data coding and quotation. Declining to respond to open-ended items did not affect participants' completion of quantitative follow-up assessments.

Statistical Analysis

Quantitative estimates in this analysis were obtained from the study dataset and cross-checked against its reported tables. Continuous variables were summarized as mean ± standard deviation and categorical variables as counts and percentages. All tests were two-sided. The primary analysis used all available repeated observations from the randomized cohort; no additional per-protocol or multiple-imputation results were generated in this analysis. The software and version, prespecified sensitivity analyses, and missing-data assumptions were documented in the empirical study.

The primary outcome used a linear mixed-effects model with fixed effects for group, categorical time, group-by-time interaction, and prior SP experience, plus a participant-specific random intercept. The overall interaction was the principal test. Time-specific adjusted differences in change, 95% confidence intervals, P values, and standardized change effect sizes were reported. The same model structure was applied to the four performance domains and self-efficacy.

Case-information errors and role-maintenance deviations used negative binomial generalized estimating equations (GEE) clustered at the SP level[32]. Anonymous undergraduate ratings used Gaussian GEE with ratings clustered within SPs. Inter-rater agreement was reported as a two-way random-effects, absolute-agreement ICC(2,1), and internal consistency as Cronbach's α. The Holm procedure was applied to the five overall interaction tests for the four performance domains and self-efficacy; count outcomes and undergraduate ratings were exploratory. Mixed models used the available observations under a missing-at-random assumption. Further sensitivity analyses were conducted in this empirical study[31].

Training satisfaction and burden scores were summarized descriptively. Their open-ended responses formed a separate qualitative component; numeric ratings alone were not treated as qualitative evidence.

Qualitative component

At T1, all questionnaire completers answered two optional open-ended items: "Which parts of training were helpful or unhelpful for your SP performance? Please give a specific example" and "What remained difficult, and what changes or further support would you suggest?" Responses were written in Chinese without a word limit. Non-response cases were recorded separately and were not coded as absence of a theme. A total of 63 participants completed the T1 questionnaires, among which 59 provided valid substantive feedback and four submitted blank responses.

For empirical analysis, two researchers independently read all responses, assigned descriptive codes, and compared their interpretations. Related codes were grouped into a set of consolidated themes using a case-by-code matrix[33]. The initial sensitizing analytical domains covered case understanding, useful learning activities, and implementation difficulties, while emergent viewpoints and conflicting perspectives were fully retained. All analytical decisions were systematically documented and linked to the original textual responses. Representative Chinese quotations were translated into English and strictly cross-checked against the original versions. All coding and thematic

procedures were formally implemented throughout the analysis; coder agreement was verified and thematic saturation was achieved in the final dataset.

Following the completion of quantitative and qualitative data summarization, a joint display was constructed to systematically align quantitative outcomes with qualitative themes[34]. The integrated interpretation identified qualitative evidence that supported, supplemented, or contextualized quantitative patterns, as well as existing inconsistencies and study limitations. Each respondent was counted only once per identified theme, irrespective of the total number of individual comments. Theme frequencies were descriptively analyzed within the respondent subset, and no statistical tests were performed to compare between-group prevalence differences. Given that open-ended feedback was solely collected at the T1 time point, the qualitative findings alone could not fully explain the sustained intervention effects observed at T2 or confirm the causal mechanism of the Combi-SP training intervention.

Results

Participant flow

A total of 71 candidate SPs entered screening. Seven did not proceed to randomization (four did not meet the inclusion criteria and three declined participation), leaving 64 SPs who were randomized to the Combi-SP group (n = 32) or conventional training group (n = 32). At T1, 32 participants in the Combi-SP group and 31 in the conventional group completed the assessment; at T2, 30 participants in each group completed the assessment.

Baseline characteristics

The baseline data showed overall balance between groups in age, sex, educational background, duration of SP work, number of previous encounters, and prior experience with rating or feedback. Absolute standardized mean differences ranged from 0.04 to 0.27. Baseline objective SP performance scores were 63.0 ± 6.1 in the Combi-SP group

and 62.0 ± 8.9 in the conventional group (SMD = 0.13). Baseline characteristics are shown in **Table 3**.

Table 3 Baseline characteristics of the two SP groups

Variable	Combi-SP group (n=32)	Conventional group (n=32)	Standardized difference (SMD)
Age (years)	36.2 ± 7.9	36.6 ± 8.9	-0.04
Female, n (%)	20 (62.5%)	24 (75.0%)	-0.27
Bachelor's degree or above, n (%)	26 (81.2%)	24 (75.0%)	0.15
Duration of SP work (months)	14.7 ± 11.6	16.3 ± 13.9	-0.13
Previous SP encounters	9.2 ± 6.7	8.9 ± 7.2	0.04
Prior student-rating experience, n (%)	11 (34.4%)	12 (37.5%)	-0.07
Prior student-feedback experience, n (%)	14 (43.8%)	13 (40.6%)	0.06
Baseline objective SP performance score	63.0 ± 6.1	62.0 ± 8.9	0.13

Primary outcome

The linear mixed-effects model showed a significant group-by-time interaction for the objective SP performance score (Wald $\chi^2(2) = 17.63$, $P < 0.001$). In the Combi-SP group, the mean score increased from 63.0 ± 6.1 at T0 to 72.9 ± 7.7 at T1 and was 71.4 ± 8.6 at T2. Corresponding scores in the conventional group were 62.0 ± 8.9, 67.1 ± 8.9, and 65.9 ± 9.1. Compared with conventional training, the adjusted between-group difference in change was 4.71 points from T0 to T1 (95% CI, 2.25–7.16; $P < 0.001$; Cohen's $d = 0.85$) and 4.43 points from T0 to T2 (95% CI, 1.93–6.93; $P < 0.001$; $d = 0.72$). The trajectories of objective SP performance scores over time are illustrated in **Fig. 3** and **Fig. 4**.

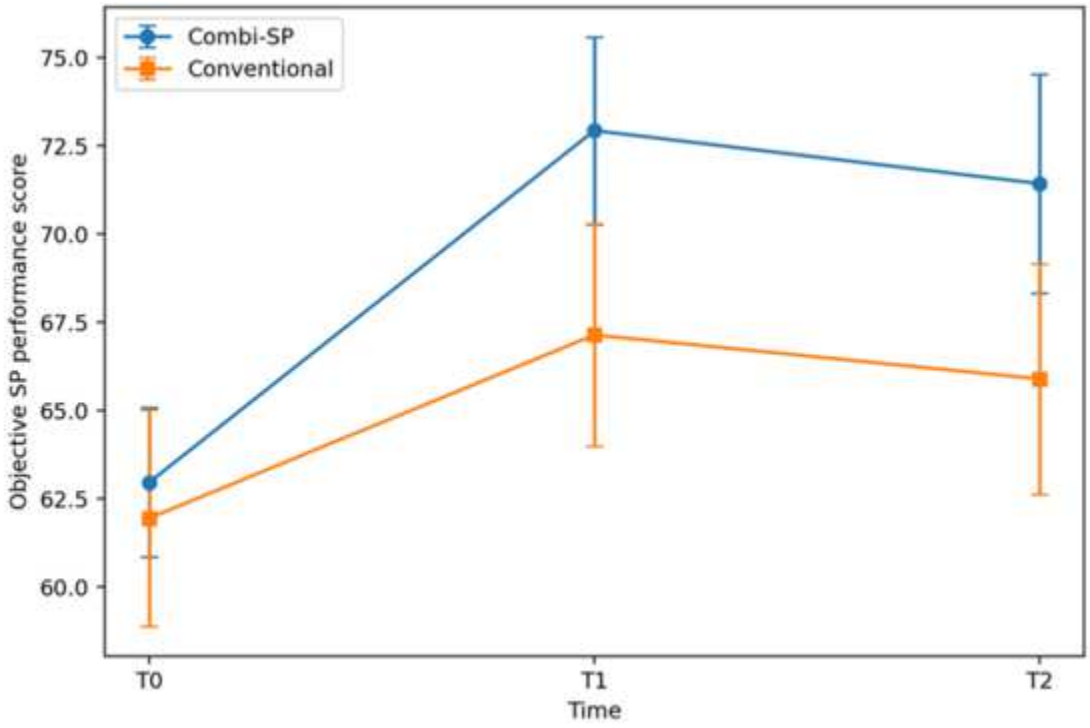


Fig. 3 Objective SP performance score over time. SP performance scores in the Combi-SP and conventional training groups at baseline (T0), immediately after training (T1), and 6-week follow-up (T2). Points represented group means, and error bars indicated 95% confidence intervals.

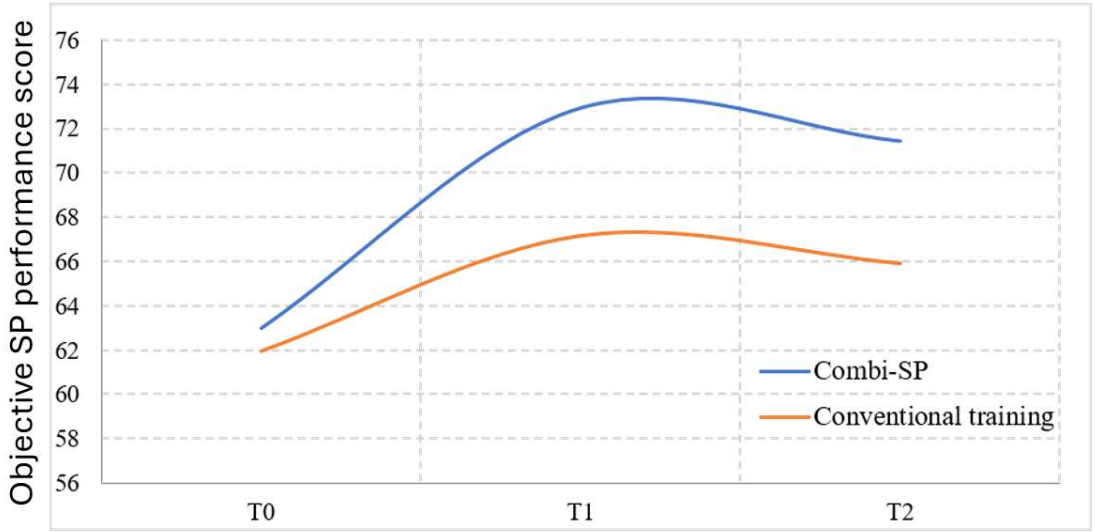


Fig. 4 Trajectory of objective SP performance score over time.

Secondary outcomes

All four objective performance domains showed group-by-time patterns favoring Combi-SP and remained statistically significant after Holm adjustment. SP self-efficacy increased from 47.8 ± 6.0 at T0 to 54.6 ± 6.3 at T1 in the Combi-SP group, compared with an increase from 48.0 ± 5.2 to 50.3 ± 6.3 in the conventional group. The total number of case-information errors decreased from 1.84 ± 1.57 at T0 to 0.62 ± 0.83 at T1 in the Combi-SP group and from 2.56 ± 1.64 to 1.84 ± 1.39 in the conventional group. In a negative binomial GEE model[32], the T1 group-by-time incidence rate ratio (IRR) was 0.49 (95% CI, 0.25–0.95; P = 0.034), but the overall interaction for this outcome was P = 0.056. The overall interaction for role-maintenance deviations was P = 0.126. Accordingly, the overall group-by-time interactions for these two count outcomes did not reach statistical significance. Clustered GEE analysis of anonymous undergraduate experience ratings showed that the Combi-SP group scored 0.20 points higher than the conventional group at T1 (95% CI, 0.14–0.26; P < 0.001). The comparison of primary and secondary outcomes are shown in **Table 4**. And the number of case-information errors is shown in **Fig. 5.**

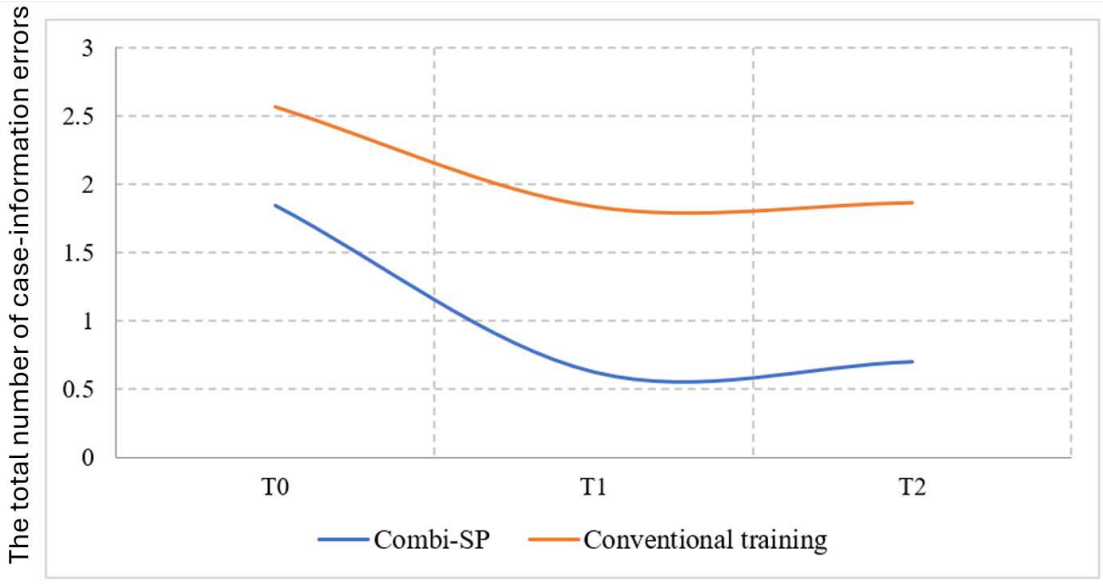


Fig. 5 Case-Information Errors Over Time. Case-information errors in the Combi-SP and conventional training groups at baseline (T0), immediately after training (T1), and

6-week follow-up (T2). Lines represented group means, with lower values indicating fewer errors.

Table 4 Comparison of primary and secondary outcomes

Outcome	Combi-SP T0	Control T0	Difference in change/IRR (95% CI)	P value
Objective SP performance score	T0 63.0±6.1 T1 72.9±7.7 T2 71.4±8.6	T0 62.0±8.9 T1 67.1±8.9 T2 65.9±9.1	T1 4.71 (2.25, 7.16) T2 4.43 (1.93, 6.93)	<0.001
Role authenticity	T0 15.7±1.5 T1 18.2±2.0 T2 18.0±2.1	T0 15.6±2.5 T1 17.0±2.3 T2 16.6±2.3	T1 1.10 (0.41, 1.79) T2 1.25 (0.55, 1.95)	0.002
Case-information accuracy	T0 18.8±1.8 T1 21.9±2.3 T2 21.3±2.7	T0 18.2±2.6 T1 19.9±2.8 T2 19.5±2.7	T1 1.43 (0.65, 2.20) T2 1.27 (0.48, 2.05)	0.002
Consistency and professional behavior	T0 12.7±1.4 T1 14.6±1.6 T2 14.2±1.7	T0 12.5±1.8 T1 13.5±1.9 T2 13.2±1.9	T1 0.97 (0.41, 1.54) T2 0.82 (0.25, 1.39)	0.002
Assessment and feedback skills	T0 15.8±1.5 T1 18.3±2.0 T2 17.9±2.2	T0 15.5±2.2 T1 16.8±2.3 T2 16.5±2.4	T1 1.20 (0.54, 1.87) T2 1.09 (0.41, 1.77)	0.002
Case-information error count	T0 1.84±1.57 T1 0.62±0.83 T2 0.70±0.84	T0 2.56±1.64 T1 1.84±1.39 T2 1.87±1.74	T1 IRR 0.49 (0.25, 0.95) T2 IRR 0.53 (0.27, 1.03)	0.056
SP self-efficacy	T0 47.8±6.0 T1 54.6±6.3 T2 53.7±6.3	T0 48.0±5.2 T1 50.3±6.3 T2 49.4±6.0	T1 4.57 (3.05, 6.10) T2 4.18 (2.63, 5.73)	<0.001
Role-maintenance deviation count	T0 0.78±0.91 T1 0.34±0.60 T2 0.33±0.61	T0 1.12±1.31 T1 1.23±1.20 T2 1.03±1.10	T1 IRR 0.40 (0.17, 0.99) T2 IRR 0.46 (0.17, 1.29)	0.126

Values are mean ± SD. Quantitative n: T0, 32 per group; T1, 32 Combi-SP and 31 conventional training; T2, 30 per group. P values for the four performance domains and self-efficacy are Holm-adjusted overall interaction tests. Count-outcome P values are unadjusted overall interaction tests. IRRs describe group-by-time interactions rather than a simple between-group rate ratio.

Reliability of the assessment instrument and inter-rater agreement

In the T0 data, the total Objective SP Performance Rating Form had a Cronbach's α of 0.98. The ICC(2,1) for the total score between the two raters was 0.92 (bootstrap 95% CI, 0.87–0.95), and domain-level ICCs ranged from 0.78 to 0.87. The reliability of the assessment instrument and inter-rater agreement are shown in **Table 5**.

Table 5 Reliability of the assessment instrument and inter-rater agreement

Instrument/domain	No. of items	Internal consistency	ICC(2,1)	95% CI
Total scale	20	α=0.98	ICC=0.92	0.87-0.95
Role authenticity	5	α=0.94	ICC=0.81	0.70-0.88
Case-information accuracy	6	α=0.93	ICC=0.87	0.77-0.92
Consistency and role maintenance	4	α=0.93	ICC=0.86	0.73-0.93
Assessment and feedback skills	5	α=0.93	ICC=0.78	0.65-0.87

Qualitative findings

The 59 responding cases contributed 118 short response units to the two open-ended items. Content analysis organized these examples into three themes. In addition, we provided a schematic diagram of the mixed-methods study design, see **Fig. 6**.

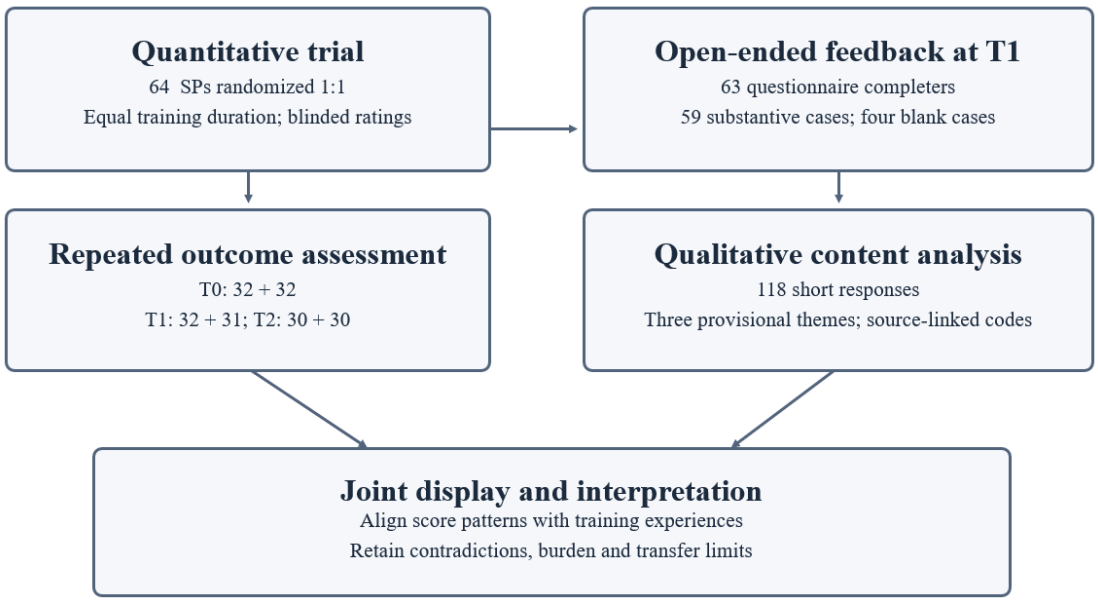


Fig. 6 Embedded mixed-methods design illustrated. The quantitative trial provides repeated performance estimates, while optional open-ended T1 feedback supplies three provisional themes.

Understanding the case and controlling disclosure

The examples distinguished understanding the patient story from memorizing isolated lines. Combi-SP comments described organizing information and deciding when to disclose it; conventional-training comments also valued familiarity with the script but highlighted uncertainty about new questions. A Combi-SP response stated: "Putting the patient story in order helped me remember what to say first and what to disclose only when asked." Remaining hesitation about persistent probing prevented an interpretation that script errors had been eliminated.

Specific feedback and targeted practice

Combi-SP examples emphasized observable targets identified through video or peer feedback and then practiced in another encounter. Conventional-training examples described helpful real-time correction alongside uncertainty when advice was general. A Combi-SP response stated: "A peer noticed that I kept looking at the rating sheet. Practicing with that specific target was more useful than general praise." Other examples described discomfort watching a recording or difficulty applying a learned approach to an unfamiliar case.

Training demands and the need for refresher support

Both groups' examples included time or emotional demands and a need to rehearse after a gap. A Combi-SP response stated: "I still needed another rehearsal for a new case rather than relying only on the previous training." These comments added practical considerations to the burden ratings. As feedback was assigned to T1, requests for later support were anticipated needs rather than observed evidence of skill decline during follow-up.

Integration of quantitative and qualitative results

The joint display in **Table 6** illustrates how the two components address related questions.

Table 6 Joint display of the quantitative and qualitative findings

Focus	Quantitative finding	Qualitative illustration	Integrated interpretation
Case understanding and disclosure	Case-information accuracy: adjusted T1 change difference 1.43 points (95% CI, 0.65–2.20). Error-count overall interaction: P = 0.056.	Story organization and disclosure boundaries. Mentioned by 15/30 Combi-SP and 20/29 conventional responders, including critical accounts.	Accounts add a possible explanation for score improvement, but persistent boundary uncertainty and non-significant error interactions limit conclusions about error reduction.
Specific feedback and practice	Objective performance: T1 difference in change 4.71 points (95% CI, 2.25–7.16). Assessment/feedback domain: 1.20 (0.54–1.87).	Specific correction targets and rehearsal. Mentioned by 15/30 and 14/29 respondents. Combi-SP examples emphasize video/peer feedback; limitations also remain.	The accounts illustrate how practice could support performance. T1 feedback does not establish causal mediation or explain six-week retention.
Training demands and ongoing support	T1 satisfaction: 4.34 ± 0.28 versus 4.05 ± 0.16. Burden: 2.27 ± 0.32 versus 2.34 ± 0.25. No burden significance test is supplied.	Time/emotional demands and refresher needs. Mentioned by 20/30 and 19/29 respondents.	Specific demands complement the averages and inform implementation. Similar burden means do not prove absence of burden or psychological safety.

Discussion

Main findings

We found that Combi-SP produced a larger immediate improvement in the objective SP performance score than conventional training, and this advantage remained at the 6-week follow-up. The four objective domains and SP self-efficacy changed in the same direction, suggesting that structured script learning, video-assisted

deliberate practice, peer calibration, and behavior-focused feedback may act together to improve SP performance. Importantly, although case-information errors showed a favorable IRR at T1, the overall group-by-time tests did not cross the conventional threshold for statistical significance.

Comparison with previous studies

Previous studies have shown that self-assessment, peer or faculty feedback, and repeated rehearsal can improve SP performance[14], that consistency-focused training can strengthen the reliability of SP ratings[15], and that multicenter calibration can improve reproducibility when different SPs portray the same case[16]. Building on this evidence, the present study brought case-information accuracy, consistency and professional behavior, and assessment and feedback skills into a single 20-item objective evaluation framework. It also used fixed standardized interviewers, blinded video rating, and repeated measurements to reduce measurement bias. A qualitative component can extend these score-based evaluations by asking SPs which activities helped, which did not, and what remained difficult.

Educational implications

SP program development should extend beyond recruitment and script memorization. A continuous quality-assurance pathway should include recruitment, basic training, case-specific training, performance assessment, rating and feedback training, periodic calibration, and refresher training[1,7,8]. The present design was particularly suitable when researchers could not directly intervene in undergraduate grades: the intervention was delivered to SPs, and the primary outcome was SP performance, while students contributed only non-graded experiential ratings. This maintained alignment among the intervention, mechanism, and outcome. Blinded rating by Office of Academic Affairs faculty may also embed research evaluation into routine SP quality management; however, study-period scores should not be used directly for dismissal, because doing so could increase participation pressure and measurement reactivity.

Strengths and limitations

Strengths of the study design included random allocation, an equal-duration control condition, fixed standardized interviewers, blinded video rating, repeated measurements. Limitations included the modest single-center sample, the need for external validation of the investigator-developed assessment instrument, residual difficulty differences across parallel cases, potential contamination between SPs, and the inability of a 6-week follow-up to establish long-term workplace performance. The results also suggested that statistical power for low-frequency count outcomes may be lower than that for continuous performance scores[29].

Conclusions

In this dataset, Combi-SP produced greater improvement in objective SP performance than conventional training, and part of the advantage was retained at 6 weeks. Evidence was stronger for continuous performance scores than for low-frequency error-count outcomes.

List of abbreviations

ASPE: Association of Standardized Patient Educators; CI: Confidence interval; Combi-SP: Competency-oriented multimodal standardized patient training program; CVI: Content validity index; GEE: Generalized estimating equations; ICC: Intraclass correlation coefficient; ICMJE: International Committee of Medical Journal Editors; IRR: Incidence rate ratio; MaSP: Maastricht Assessment of Simulated Patients; OSCE: Objective structured clinical examination; SMD: Standardized mean difference; SP: Standardized patient; SPSS: Statistical Package for the Social Sciences; T0: Baseline assessment; T1: Immediate post-training assessment; T2: Six-week post-training follow-up assessment;

Declarations

Ethics approval and consent to participate

This study was approved by the Medical Research Ethics Committee of The Seventh Affiliated Hospital, Sun Yat-sen University (Shenzhen) (Approval No. KY-

2026-313-01). The research was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants prior to enrollment.

Consent for publication

Written informed consent was obtained from all participants prior to their participation in the study.

Availability of data and materials

All data supporting the conclusions of this research are available upon reasonable request from the corresponding author.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by Guangdong University Quality Assurance Project (Grant No. 78000-12263012).

Authors' Contributions

All authors have met the ICMJE criteria for authorship, have read and approved the final manuscript, and have made the following contributions: Conceptualization: MingYue Zhou; Methodology: Qian Chen, Xin Shu; Formal analysis and investigation: Xin Shu, Mingqing Zhao; Writing - original draft preparation: MingYue Zhou, Yawen Zhang; Writing - review and editing: MingYue Zhou, Yawen Zhang, Zhenning Liang; Resources: Qian Chen, Jingfang Luo, Zhenning Liang; Supervision: Jiancheng Wang.

Acknowledgements

The authors would like to thank all study participants.

Corresponding author

Correspondence to Jiancheng Wang.

References

1. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul (Lond). 2017;2:10. doi:10.1186/s41077-017-0043-4.

2. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach. 2009;31(6):477-486. doi:10.1080/01421590903002821.

3. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med. 1993;68(6):443-451. doi:10.1097/00001888-199306000-00002.

4. Siemerkus J, Petrescu AS, Köchli L, Stephan KE, Schmidt H. Using standardized patients for undergraduate clinical skills training in an introductory course to psychiatry. BMC Med Educ. 2023;23(1):159. doi:10.1186/s12909-023-04107-5.

5. Bokken L, Rethans JJ, Scherpbier AJJA, van der Vleuten CPM. Strengths and weaknesses of simulated and real patients in the teaching of skills to medical students: a review. Simul Healthc. 2008;3(3):161-169. doi:10.1097/sih.0b013e318182fc56.

6. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447-451. doi:10.1136/bmj.1.5955.447.

7. Furman GE. The role of standardized patient and trainer training in quality assurance for a high-stakes clinical skills examination. Kaohsiung J Med Sci. 2008;24(12):651-655. doi:10.1016/s1607-551x(09)70031-2.

8. Peters T, Sommer M, Fritz AH, Kursch A, Thrien C. Minimum standards and development perspectives for the use of simulated patients - a position paper of the committee for simulated patients of the German Association for Medical Education. GMS J Med Educ. 2019;36(3):Doc31. doi:10.3205/zma001239.

9. Wind LA, Van Dalen J, Muijtjens AM, Rethans JJ. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). Med Educ. 2004;38(1):39-44. doi:10.1111/j.1365-2923.2004.01686.x.

10. Bokken L, Linssen T, Scherpbier A, van der Vleuten C, Rethans JJ. Feedback by simulated patients in undergraduate medical education: a systematic review of the literature. Med Educ. 2009;43(3):202-210. doi:10.1111/j.1365-2923.2008.03268.x.

11. Howley L, Szauter K, Perkowski L, Clifton M, McNaughton N; Association of Standardized Patient Educators (ASPE). Quality of standardised patient research reports in the medical education literature: review and recommendations. Med Educ. 2008;42(4):350-358. doi:10.1111/j.1365-2923.2007.02999.x.

12. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011;306(9):978-988. doi:10.1001/jama.2011.1234.

13. Cen X, Wu P, Fang S, Ma K, Wang N, Huang J, et al. Bridging Basic and Clinical Medicine: A Combi-Method Approach to Pathophysiology Education. J Med Educ Curric Dev. 2025;12:23821205251405327. doi:10.1177/23821205251405327.

14. Perera J, Perera J, Abdullah J, Lee N. Training simulated patients: evaluation of a training approach using self-assessment and peer/tutor feedback to improve performance. BMC Med Educ. 2009;9:37. doi:10.1186/1472-6920-9-37.

15. MacLean S, Geddes F, Kelly M, Della P. Simulated patient training: Using inter-rater reliability to evaluate simulated patient consistency in nursing education. Nurse Educ Today. 2018;62:85-90. doi:10.1016/j.nedt.2017.12.024.

16. Resende KA, Cavaco AM, Luna-Leite MD, Acacio BR, Pinto NN, Neta MD, et al. Training and standardization of simulated patients for multicentre studies in clinical pharmacy education. Pharm Pract (Granada). 2020;18(4):2038. doi:10.18549/pharmpract.2020.4.2038.

17. Gonullu I, Dogan CD, Erden S, Gokmen D. A study on the standard setting, validity, and reliability of a standardized patient performance rating scale - student version. Ann Med. 2023;55(1):490-501. doi:10.1080/07853890.2023.2168744.

18. Wu ML, Chao LF, Hung YT. Training students to serve as standardized patients in an objective structured clinical examination is feasible: A mixed-methods study. Nurse Educ Pract. 2024;79:104069. doi:10.1016/j.nepr.2024.104069.

19. Sommer M, Fritz AH, Thrien C, Kursch A, Peters T. Simulated patients in medical education - a survey on the current status in Germany, Austria and Switzerland. GMS J Med Educ. 2019;36(3):Doc27. doi:10.3205/zma001235.

20. Frank JR, Snell LS, ten Cate O, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. Med Teach. 2010;32(8):638-645. doi:10.3109/0142159x.2010.501190.

21. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27(1):10-28. doi:10.1080/01421590500046924.

22. Cook DA, Hamstra SJ, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. Med Teach. 2013;35(1):e867-e898. doi:10.3109/0142159x.2012.714886.

23. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med. 2004;79(10 Suppl):S70-S81. doi:10.1097/00001888-200410001-00022.

24. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med. 2011;86(6):706-711. doi:10.1097/acm.0b013e318217e119.

25. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014;348:g1687. doi:10.1136/bmj.g1687.

26. van de Ridder JMM, Stokking KM, McGaghie WC, ten Cate OTJ. What is feedback in clinical education? Med Educ. 2008;42(2):189-197. doi:10.1111/j.1365-2923.2007.02973.x.

27. Lynn MR. Determination and quantification of content validity. Nurs Res. 1986;35(6):382-385. doi:10.1097/00006199-198611000-00017.

28. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health. 2006;29(5):489-497. doi:10.1002/nur.20147.

29. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53-55. doi:10.5116/ijme.4dfb.8dfd.

30. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012.

31. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367(14):1355-1360. doi:10.1056/nejmsr1203730.

32. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22. doi:10.1093/biomet/73.1.13.

33. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol. 2013;13:117. doi:10.1186/1471-2288-13-117.

34. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs—principles and practices. Health Serv Res. 2013;48(6 Pt 2):2134-2156. doi:10.1111/1475-6773.12117.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymeterial.docx
- StatisticalTable.xlsx