Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Reliable Assessment of Healthcare Procedures: A Comparison of On-Site and Video-Based Methods

Minghui Yi[1], Yao Zhang[2], Fei Wu[1], Shilong Zhang[3], Menghui Hu[1], Changmei Liu[1], Bin Zheng[2], Jinling Yang[1]

[1]Medical Integration and Practice Center, Cheeloo College of Medicine, Shandong University, Jinan, Shandong, People's Republic of China; [2]Surgical Simulation Research Lab, Department of Surgery, University of Alberta, Edmonton, Alberta, Canada; [3]Department of Social Medicine and Health Management, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, Shandong, People's Republic of China

Correspondence: Jinling Yang, Medical Integration and Practice Center, Cheeloo College of Medicine, Shandong University, No. 44 Wenhua Xi Road, Jinan, Shandong, People's Republic of China, Email yjlxixi@126.com

Introduction: Assessing performance is critical in simulation training programs. Traditionally, predefined evaluation forms completed by experts are used, but this approach may introduce bias. Common strategies to mitigate bias include averaging scores from multiple evaluators or using video recordings to minimize disagreements among assessors. This study compares performance scores obtained through real-time on-site observation, independent video assessment, and collaborative video assessment of venipuncture performance. Additionally, we evaluate whether combining these methods enhances scoring consistency.

Methods: Eighteen medical students were invited to perform venipuncture trials, which were evaluated in three stages. Two evaluators observed and scored each trial on-site using a predefined evaluation form. The trials were video-recorded. At 12 weeks post-training, the evaluators independently reviewed the videos and assigned performance scores. At 14 weeks, they collaboratively reviewed the videos and provided a joint performance score. The Intraclass Correlation Coefficient (ICC) was used to assess the consistency between evaluators.

Results: Mean scores differed significantly among the three assessment methods (P = 0.037), with independent video assessment yielding higher scores (75.3 ± 5.2) than on-site assessment (66.7 ± 6.1). Inter-rater reliability (ICC) ranged from 0.706 to 0.883, but was not statistically compared across methods.While collaborative assessment showed the highest consistency (0.889), implementing a multimodal approach would substantially increase faculty workload, as it requires scoring each student multiple times.

Conclusion: Video-based assessments, particularly collaborative review, enable detailed and repeated analysis of procedural skills, improving scoring consistency. However, the feasibility and workload implications of combining multiple methods must be considered before implementation in training programs.

Keywords: simulation, performance assessment, video-based assessment, collaborative scoring, inter-rater reliability

Background

High-quality healthcare services require strong clinical judgment and excellent human performance, particularly in the delivery of therapeutic health procedures.[1] Clinical skills training is a crucial component of health education, ensuring that trainees develop the necessary competencies over their long course of healthcare study.[2] Simulation has been widely adopted as an effective method for skill training, as it provides a safe learning environment with sufficient realism and reasonable teaching costs.[3] While many educators focus on designing simulation-based training curricula and evaluating learning outcomes, an important yet often overlooked research area is the assessment of human performance during simulation-based skill practice.[4]

Currently, most performance assessments are conducted by content experts using pre-designed evaluation forms while observing trainees in real-time.[5] This on-site assessment method is quick and reliable, leveraging the experts' instinctive

https://doi.org/10.2147/AMEP.S592969
Received: 31 December 2025
Accepted: 1 April 2026
Published: 13 April 2026

judgment.[6] However, it has notable limitations. Performance scores may be influenced by external factors in the training environment and internal biases of the evaluators.[7,8] Additionally, verifying the consistency of assessments, whether across different trainees or between multiple evaluators assessing the same trainee remains a challenge.[9] To mitigate these limitations, clinical skills training often involves multiple evaluators, averaging their scores to reduce individual biases.[10]

Another approach to improving skill evaluation reliability is video-based assessment, where training sessions are recorded, and trainees' performance is scored later.[11] By allowing evaluators to review, replay, and analyze videos at their convenience,[12] video-based assessments offer greater flexibility and may reduce variability through multiple viewings, thereby providing opportunities for more deliberate and reflective judgment. Multiple evaluators can assess the same video independently or collaboratively, discussing discrepancies and may lead to more consistent judgments. From a social constructivist perspective, this process can be understood as evaluators co-constructing shared understandings of performance standards through discussion and negotiation, which aligns with consensus-based approaches to assessment reliability.[13] Furthermore, secondary analyses of these recordings can help identify behavioral changes in trainees over time.

Although video-based assessments require more time, they are increasingly used in health education due to their numerous advantages.[14] However, while previous studies have examined the reliability of video-based assessment compared to on-site observation,[15,16] few have directly compared the reliability of independent video review and collaborative video review within the same cohort of trainees and evaluators. In particular, the extent to which collaborative discussion among raters enhances scoring consistency remains underexplored. As clinical training programs seek to implement fair and efficient assessment methods, evidence is needed to guide the selection of approaches that balance reliability, feasibility, and educational impact. This study therefore aimed to compare inter-rater reliability across three assessment modalities-on-site, independent video, and collaborative video-for evaluating a procedural clinical skill (venipuncture), and to examine differences in mean scores across these conditions.

To address the aforementioned questions, we conducted a controlled laboratory study in which a group of health students performed a blood collection procedure via venipuncture (phlebotomy) in a simulation setting. Venipuncture was chosen for assessment because it is a fundamental clinical skill that medical students must master and a mandatory component of the Licensing Examination for Medical Practitioners in China. Like other procedural tasks, evaluating venipuncture requires a step-by-step assessment of various stages, including preoperative preparation, patient positioning, skin disinfection, venipuncture and blood collection, sample handling, and the organization of materials.[17]

The primary objective of our study was to determine whether video-based assessment offers advantages over on-site scoring by comparing the consistency of evaluations across different assessment methods. Assessment quality was measured by the consistency of scores given by two expert evaluators across multiple assessment modalities, including real-time on-site evaluation and video-based evaluation conducted both independently and collaboratively. Additionally, we aimed to explore whether combining on-site and video-based assessments could improve overall assessment reliability.

Our research hypotheses are as follows:

1. Video-based assessment of the venipuncture procedure will demonstrate higher consistency compared to on-site assessment.
2. Greater consistency will be achieved when combining different assessment methods, such as: on-site assessment combined with independent video-based assessment by two evaluators, on-site assessment combined with collaborative video-based assessment, and independent and collaborative video-based assessments combined. These combinations will yield more consistent results than on-site assessment alone.
3. When all three assessment methods (on-site, independent video-based, and collaborative video-based) are combined, the evaluation results will demonstrate the highest level of consistency compared to using only on-site assessment.

Methods

Participants

Participants were recruited from second- and third-year clinical medical students at Shandong University. Eligibility criteria required that they had not received prior specialized training in venipuncture. Our study adhered to the principles

outlined in the Declaration of Helsinki. The project was approval from the Ethics Committee of the School of Clinical Medicine, Shandong University (SDULCLL2022-20). Before participation, all students signed an informed consent form, and their study schedules were arranged accordingly.

Task and Procedure

All participants completed a 90-minute training course, which included a PowerPoint presentation on venipuncture-related knowledge, a demonstration video of the procedure, and a hands-on demonstration. In the subsequent skills training session, each participant engaged in supervised practice sessions on a specialized venipuncture simulation model (Shandong Yifutai Manufacturer, Model SY/H008, Jinan, China). Over the course of one week, each participant completed practice trials on four randomly selected days, performing three trials per practice day. These practice sessions were for skill acquisition only and were not included in the analysis.

On the seventh day, each participant completed one formal venipuncture test using the same simulation model. This test trial was video recorded and served as the trial for all subsequent score evaluations. The test was assessed under three conditions: (1) on-site evaluation by two independent assessors immediately after the test; (2) independent video-based evaluation by the same two assessors 12 weeks later; and (3) collaborative video-based evaluation, where both assessors reviewed and discussed the recordings together 14 weeks later. A total of 18 tests (one per participant) were included in the analysis.

All 18 participants were assessed under all three conditions. Thus, each participant served as their own control, and all comparisons are within-subjects.

Evaluators and Scoring Criteria

Participants' performance was evaluated by two expert assessors from the Medical Simulation Center of Shandong University. Both evaluators had over 15 years of experience teaching medical skills using simulation and extensive expertise in performance assessment. Additionally, they were involved in developing the evaluation criteria for venipuncture.

The study was designed to ensure that evaluators remained blinded to student identity throughout all assessment phases. The evaluators had no prior knowledge of the students and did not interact with them during the assessment process. For on-site assessments, students were identified only by code numbers. For video-based assessments, although student faces were visible in the recordings, the evaluators did not have access to any identifying information linking the faces to student names or codes.

Performance assessment was based on four key steps of the venipuncture procedure: including pre-operation preparation, site inspection, venous blood collection, and post-puncture handling. Due to varying complexity levels of each step, different point values were assigned accordingly. Specifically, pre-operation preparation has 16 points, site inspection has16 points, venous blood collection has 40 points, and post-puncture handling has 18 points. In addition to the 90 points allocated for procedural steps, an additional 10 points were awarded for overall professionalism, which included patient care and communication skills.

The total score was calculated out of 100 points, with higher scores indicating better performance. The scoring system was developed by five clinical experts and has been nationally adopted for assessing venipuncture performance. The scoring sheets used for assessment are provided in Supplementary Table 1.

On-Site Assessment

Before the simulation training, the two evaluators spent 20 minutes reviewing and discussing the scoring criteria to ensure a shared understanding of how to assign points for each step of the procedure. Participants entered the training room in a randomly determined order to perform the venipuncture procedure. During the assessment, the two evaluators independently evaluated the participants' performance simultaneously, without any communication or discussion between themselves or with the participants.

Video-Based Assessment

All venipuncture procedures were recorded using a video recording system installed at the Medical Simulation Center. The system included three cameras (2DE22041W-D3, Hikvision, Hangzhou, China): one positioned at the entrance of the training room and two mounted on the left and right walls, approximately 2.2 meters above the ground. These cameras continuously captured the participants' movements and audio throughout the procedure.

To ensure adequate visualization of fine motor skills essential for venipuncture assessment, a pilot test was conducted prior to data collection. An assistant performed the venipuncture procedure while the research team verified that all critical steps (eg., needle insertion angle, flashback observation) were clearly visible from at least one camera angle. The multi-angle setup allowed for complementary views, ensuring that any details potentially missed by one camera were captured by another. The cameras recorded at a resolution of 1920×1080 pixels, which provided sufficient clarity for evaluating procedural details.

At the 12th week after the training session, the two evaluators were called back to independently watch and assess the recorded venipuncture performances. They were not allowed to discuss or exchange opinions during this evaluation. The 12-week gap was chosen to minimize the evaluators' recollection of the scores they assigned during the on-site assessment.

At week 14 post-training, the evaluators were invited back to jointly review and score the videos. Unlike the independent assessments, this time they discussed their observations, exchanged opinions, and assigned a single joint score for each selected venipuncture trial.

Statistical Analysis

Statistical analyses were conducted using SPSS 21.0. The scores assigned by the two evaluators for different participants were reported as mean ± standard deviation. To compare mean scores across the three assessment methods (on-site, independent video, and collaborative video), a one-way repeated measures analysis of variance (ANOVA) was conducted. This approach was chosen because the same 18 participants were assessed under all three conditions, making the data dependent (within-subjects design). Prior to analysis, the scores from the two evaluators were averaged to generate a single representative score for each assessment method (on-site mean, independent video mean, and collaborative video score). Post-hoc pairwise comparisons were performed using Bonferroni correction to control for Type I error. A P-value was calculated to determine the significance of differences among averages, with results considered statistically significant when $P < 0.05$.

The Intraclass Correlation Coefficient (ICC) was used to assess the consistency between the two evaluators. Based on the methods described by Shrout & Fleiss and McGraw & Wong, a two-way random-effects model was applied to measure absolute agreement among different evaluators.[18,19] This study specifically reported absolute agreement, as the scores assigned to different participants within the same group were not expected to be correlated in an additive manner.

A two-way random-effects model was chosen because each participant was evaluated by the same two independent evaluators, who were randomly selected from a pool of evaluators. The basic model is:

$$y_{ij} = \mu + r_i + c_j + \in_{ij}$$

Where y_{ij} is the score given by the j-th evaluator to the i-th target, μ is the mean score, r_i is the target random effect, c_j is the evaluator random effect, where $\in_{ij}$ is the random error.

Absolute agreement among different participants (AA-ICC):

$$\mathrm{ICC}\,(\mathrm{A},\,1) = \frac{\sigma_{\mathrm{r}}^2}{\sigma_{\mathrm{r}}^2 + \sigma_{\mathrm{c}}^2 + (\sigma_{\mathrm{rc}}^2 + \sigma_{\in}^2)}$$

Absolute agreement among different evaluators (AA-ICC):

$$\mathrm{ICC}\,(\mathrm{A},\,\mathrm{k}) = \frac{\sigma_{\mathrm{r}}^2}{\sigma_{\mathrm{r}}^2 + (\sigma_{\mathrm{c}}^2 + \sigma_{\mathrm{rc}}^2 + \sigma_{\in}^2)/\mathrm{k}}$$

According to Landis and Koch's guidelines, an ICC value greater than 0.80 indicates very strong agreement, a value between 0.60 and 0.80 indicates strong agreement, a value between 0.40 and 0.60 indicates moderate agreement, a value between 0.20 and 0.40 indicates fair agreement, and a value less than 0.20 indicates poor agreement.[20]

Results

Participants Performance Scores Given by Different Assessment Methods

A one-way repeated measures ANOVA revealed a significant main effect of assessment method on mean scores ($F(2, 34) = 44.338$, $p < 0.001$, partial $\eta^2 = 0.723$). Post-hoc comparisons indicated that all three methods differed significantly from each other (all $p < 0.05$). Detailed descriptive statistics and pairwise comparisons are presented in Table 1.

Consistency Between the Two Evaluators Using a Single Evaluation Method

The consistency between the two evaluators when using a single evaluation method is presented in Table 2. The intraclass correlation coefficient (ICC) for on-site scoring was 0.778, while the ICC for independent video-based scoring was 0.706.

Consistency Between the Two Evaluators Using Combinations of Two Evaluation Methods

When combining two evaluation methods, the consistency between the two evaluators is shown in Table 3. The ICC for the combination of on-site scoring and independent video-based scoring was 0.835, for the combination of on-site scoring and collaborative video-based scoring was 0.883, and for the combination of independent video-based scoring and collaborative video-based scoring was 0.841.

Consistency Between the Two Evaluators Using Combinations of Three Evaluation Methods

Consistency between the two evaluators when combining all three evaluation methods is shown in Table 4. The ICC for the combination of on-site scoring, independent video-based scoring, and collaborative video-based scoring was 0.889.

Table 1 Mean Scores and Pairwise Comparisons for the Three Assessment Methods

Assessment Method	Mean ± SD	Pairwise Comparisons	Mean Difference (95% CI)	Adjusted p-value
On-site	66.72 ± 10.67	vs Independent video	8.56 (5.96 to 11.15)	<0.001
		vs Collaborative video	3.06 (0.31 to 5.80)	0.027
Independent video	75.28 ± 9.07	vs On-site	8.56 (5.96 to 11.15)	<0.001
		vs Collaborative video	5.50 (3.58 to 7.42)	<0.001
Collaborative video	69.79 ± 9.64	vs On-site	3.06 (0.31 to 5.80)	0.027
		vs Independent video	5.50 (3.58 to 7.42)	<0.001

Note: P-values are adjusted using Bonferroni correction for multiple comparisons. Repeated measures ANOVA showed a significant main effect of assessment method ($F(2, 34) = 44.338$, $p < 0.001$, partial $\eta^2 = 0.723$). n = 18 for all groups.

Table 2 Consistency Between the Two Evaluators Using a Single Evaluation Method

	On-Site		Independent Video	
	ICC	95% CI	ICC	95% CI
Individual	0.637	(−0.060, 0.884)	0.546	(−0.034, 0.873)
Average	0.778	(−0.127, 0.938)	0.706	(−0.071, 0.932)

Note: Individual ICC refers to single-rater reliability; Average ICC refers to the reliability of the mean of two raters. 95% confidence intervals are shown in parentheses. n = 18 for all groups.

https://doi.org/10.2147/AMEP.S592969

Table 3 Consistency Between the Two Evaluators Using Combinations of Two Evaluation Methods

	On-Site + Independent Video		On-Site + Collaborative Video		Independent Video + Collaborative Video	
	ICC	95% CI	ICC	95% CI	ICC	95% CI
Individual	0.558	(0.135, 0.817)	0.716	(0.300, 0.893)	0.639	(0.055, 0.883)
Average	0.835	(0.384, 0.947)	0.883	(0.563, 0.961)	0.841	(0.149, 0.958)

Note: Individual ICC refers to single-rater reliability; Average ICC refers to the reliability of the mean of two raters. 95% confidence intervals are shown in parentheses. n = 18 for all groups.

Table 4 Consistency Between the Two Evaluators Using Combinations of Three Evaluation Methods

	On-Site + Independent Video + Collaborative Video	
	ICC	95% CI
Individual	0.616	(0.242, 0.838)
Average	0.889	(0.614, 0.963)

Note: Individual ICC refers to single-rater reliability; Average ICC refers to the reliability of the mean of two raters. 95% confidence intervals are shown in parentheses. n = 18 for all groups.

It should be noted that the ICC estimates presented in Tables 2–4 are reported descriptively with 95% confidence intervals; no statistical comparisons were performed across different methods or combinations. The confidence intervals show substantial overlap across conditions (eg., on-site Average ICC: 0.778, 95% CI [−0.127, 0.938]; independent video Average ICC: 0.706, 95% CI [−0.071, 0.932]), indicating that apparent differences in reliability should be interpreted cautiously, as they may not represent statistically significant differences.

Discussion

Currently, on-site assessment on human performance is widely used in clinical skills setting, where two evaluators independently score the same performance, and the average score is taken as the final assessment result.[21] While this method can, to some extent, reflect the operator's true proficiency, it has limitations. The on-site assessment environment is often noisy, the operation time is short, and the steps are performed rapidly, making it difficult for evaluators to capture all details.[22] Additionally, scoring can be influenced by subjective factors such as the evaluators' experience, leading to inconsistencies and reduced reliability.[23,24]

The video-based assessment allows evaluators to review the videos subsequently, helping to compensate for the shortcomings of on-site scoring and optimizing the assessment process through the combination of different evaluation methods.[25] The study results indicate a significant difference in the average scores among on-site scoring, independent video-based scoring and collaborative video-based scoring ($P < 0.05$). This discrepancy may be due to the fast-paced nature of on-site assessments, where certain scoring points can be overlooked.[26] In contrast, video-based assessment allows evaluators to use slow playback and rewind functions for a more detailed review.[27] When using video-based collaborative assessment, discussion between the two evaluators helps clarify discrepancies and reach consensus, thereby enhancing scoring consistency.

When a single evaluation method was used, the consistency of on-site scoring between the two evaluators was 0.778, indicating strong agreement, while the consistency of independent video-based scoring was 0.706, also reflecting strong agreement. The ICC value for on-site scoring was slightly higher than that of independent video-based scoring, a result that does not fully align with our initial hypothesis. We had anticipated that video-based scoring would enhance inter-rater consistency in assessing operators' skill levels, as evaluators could replay and closely examine specific steps.[12] However, on-site evaluators shared the same environment, whereas video-based evaluations were conducted separately under different conditions.[28,29] Factors such as the way the video was viewed, the time of viewing, and the camera angle may have contributed to the slightly lower consistency in video-based scoring compared to on-site scoring.[26,30]

When combining two evaluation methods, the consistency between the two evaluators for on-site scoring and independent video-based scoring was 0.835, for on-site scoring and collaborative video-based scoring was 0.883, and for

independent video-based scoring and collaborative video-based scoring was 0.841. These results indicate that agreement between the two evaluators was very strong when using two evaluation methods and significantly better than when using a single evaluation method, which aligns with our second hypothesis. Among these combinations, the highest consistency was observed when combining on-site scoring with collaborative video-based scoring. This is likely because the evaluators watched the videos in the same environment, allowing for real-time discussion of scoring discrepancies.[30]

This finding can be understood through a social constructivist lens, where discussion and negotiation among evaluators facilitate the co-construction of shared understandings regarding performance standards.[13] By replaying the videos and reviewing controversial scoring points together, they could achieve greater inter-rater consistency in assessing the operator's skill level.[31] The potential of collaborative review to serve as a calibration exercise is also consistent with literature on rater cognition, which suggests that scoring variability often stems from differences in how evaluators interpret and apply criteria, rather than from the information available to them.[32] This interpretation is further supported by recent studies in interprofessional education, where peer assessment of video-based teamwork demonstrated good reliability, although assessor subjectivity remained a notable source of variance.

However, collaborative video-based scoring has certain limitations, as it requires evaluators to coordinate time and location for joint evaluation. Therefore, its use should be considered based on practical constraints. When combining independent and collaborative video-based scoring, assessments can be conducted without on-site evaluators, reducing the performer's stresses and potentially providing a more authentic demonstration of clinical skills.[31] This method also offers greater flexibility in scheduling for evaluators while maintaining high scoring consistency.

The consistency between the two evaluators when combining all three assessment methods, the on-site, the independent video-based scoring, and the collaborative video-based assessment was 0.889, slightly higher than the combination of any two but no significantly. This finding supports our third research hypothesis that consistency improves with multiple evaluation methods. However, using all three methods requires additional time and effort. Therefore, choosing a combination of two assessment methods maintains an optimal balance between consistency and efficiency.

The integration of video recording enhances the consistency of skill assessments and improves the quality of clinical skills training.[33] Video evidences also enhance the precision when provide feedback to performers, allowing them to review their own performance, identify errors, and learn from each other.[34] This approach fosters continuous improvement in clinical skills and competency.

While this study demonstrates differences in inter-rater reliability across assessment modalities, it is important to clarify that these findings pertain to scoring consistency, not assessment accuracy. Without an external gold standard, we cannot determine which method yields scores that more closely reflect students' true clinical competence. Rather, the contribution of this study lies in comparing how reliably different assessment formats perform under controlled conditions, which has practical implications for program evaluation and faculty training.

This study has several limitations. First, the sample size (N = 18) is relatively small and only two evaluators participated, which may limit statistical power and the generalizability of the findings. Future research should include larger and more diverse participant cohorts, involve a greater number of evaluators, and refine scoring criteria to conduct more comprehensive consistency analyses. Second, this study investigated the reliability of test scores done by evaluators on the test site versus by videos. As no external gold standard was available to validate the assessment scores against true performance, we did not examine the validity and accuracy. High inter-rater agreement may reflect shared biases rather than accurate measurement of clinical competence.

Future research should address these limitations through several approaches. Larger-scale studies with more diverse participant cohorts would enhance generalizability. Development of objective performance metrics could provide an external gold standard for validating video-based assessments. Studies examining the feasibility and workload implications of implementing collaborative assessment in real-world training programs are also needed.

Conclusion

In summary, Video-based assessments, particularly collaborative review, improve scoring consistency by enabling detailed and repeated analysis of procedural skills. However, they remain influenced by evaluator subjectivity, scoring environments, and video quality. While Selecting an appropriate combination of scoring methods can result in improved

consistency and reliability, the feasibility and workload implications of combining multiple methods must be carefully considered before implementation in training programs.

Abbreviations

ICC, Intraclass Correlation Coefficient.

Data Sharing Statement

Data will be made available upon request to health educators for legitimate research purposes.The data that support the findings of this study are available from the corresponding author, Jinling Yang, upon reasonable request.

Ethics Approval and Consent to Participate

Our study adhered to the principles outlined in the Declaration of Helsinki. Ethical approval was obtained from the Ethics Committee of the School of Clinical Medicine, Shandong University (SDULCLL2022-20), and all participants provided informed consent prior to their involvement in the study.

Consent for Publication

This manuscript has been read and approved by all authors.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research was supported by grant M2023357 from the Undergraduate Teaching Reform Research Project in Shandong Province and grant sy20242401 from the Research Program on Laboratory Construction and Management at Shandong University and grant qlyxjy-202417 from Undergraduate Teaching Reform Project of Cheoloo College of Medicine at Shandong University, China.

Disclosure

The authors declare no competing interests in this work.

References

1. Schwarze-Chintapatla A, Livingstone-Banks J, Chua J, Albury C. Is behaviour change communication guidance for general practice healthcare professionals evidence based? A systematic review. *BMC Med*. 2026;24(1):111. doi:10.1186/s12916-026-04681-7
2. Yale SH, Tekiner H, Yale ES. Eponymous clinical signs, clinical skills and the humanities. *Br J Hosp Med*. 2023;84(9):1. doi:10.12968/hmed.2023.0236
3. Bray L, Krogh TB, Østergaard D. Simulation-based training for continuing professional development within a primary care context: a systematic review. *Educ Prim Care*. 2023;34(2):64–73. doi:10.1080/14739879.2022.2161424
4. Korayem GB, Alshaya OA, Kurdi SM, et al. Simulation-based education implementation in pharmacy curriculum: a review of the current status. *Adv Med Educ Pract*. 2022;13:649–660. doi:10.2147/AMEP.S366724
5. Gilani S, Pankhania K, Aruketty M, et al. Twelve tips to organise a mock OSCE. *Med Teach*. 2022;44(1):26–31. doi:10.1080/0142159X.2021.1887465
6. Tsao Y-T, Huang T-S, Chen H-L, Chu T-S, Lee M-B, Chen -Y-Y. Examining whether Direct Observation of Procedural Skill (DOPS) is a reliable assessment tool for assessing the performance of procedural skills. *J Med Educ*. 2021;25(3):103–111.
7. Yeates P, Maluf A, McCray G, et al. Inter-school variations in the standard of examiners' graduation-level OSCE judgements. *Med Teach*. 2025;47(4):735–743. doi:10.1080/0142159X.2024.2372087
8. Yeates P, Moult A, Cope N, et al. Measuring the effect of examiner variability in a multiple-circuit objective structured clinical examination (OSCE). *Acad Med*. 2021;96(8):1189–1196. doi:10.1097/ACM.0000000000004028
9. Cade AE, Meuller N. Measuring the quality of the OSCE in a chiropractic programme: a review of metrics and recommendations. *J Chiropr Educ*. 2024;38(1):9–16. doi:10.7899/JCE-22-29

10. Touma NJ, Paco CA, MacIntyre I. Inter-observer variance of examiner scoring in urology objective structured clinical examinations. *Can Urol Assoc J*. 2024;18(4):116–119. doi:10.5489/cuaj.8571
11. Addison P, Bitner D, Carsky K, et al. Outcome prediction in bariatric surgery through video-based assessment. *Surg Endosc*. 2023;37(4):3113–3118. doi:10.1007/s00464-022-09480-8
12. Balvardi S, Semsar-Kazerooni K, Kaneva P, et al. Validity of video-based general and procedure-specific self-assessment tools for surgical trainees in laparoscopic cholecystectomy. *Surg Endosc*. 2023;37(3):2281–2289. doi:10.1007/s00464-022-09466-6
13. Liao KC, Ajjawi R, Peng CH, Jenq CC, Monrouxe LV. Striving to thrive or striving to survive: professional identity constructions of medical trainees in clinical assessment activities. *Med Educ*. 2023;57(11):1102–1116. doi:10.1111/medu.15152
14. Mitchell O, Cotton N, Leedham-Green K, Elias S, Bartholomew B. Video-assisted reflection: improving OSCE feedback. *Clin Teach*. 2021;18(4):409–416. doi:10.1111/tct.13354
15. Lund S, Navarro S, D'Angelo JD, Park YS, Rivera M. Expanded access to video-based laparoscopic skills assessments: ease, reliability, and accuracy. *J Surg Educ*. 2024;81(6):850–857. doi:10.1016/j.jsurg.2024.03.010
16. Sinha A, Nimbalkar SM, Pujara RK, et al. SimCapture app video performance assessment versus real-time instructor-based performance evaluation of undergraduates in neonatal resuscitation-an agreement study. *J Trop Pediatr*. 2024;70(6). doi:10.1093/tropej/fmae033
17. O'Herlihy N, Griffin S, Henn P, Gaffney R, Cahill MR, Gallagher AG. Validation of phlebotomy performance metrics developed as part of a proficiency-based progression initiative to mitigate wrong blood in tube. *Postgrad Med J*. 2021;97(1148):363–367. doi:10.1136/postgradmedj-2019-137254
18. Bamney A, Jashami H, Sonduru Pantangi S, et al. Examining impacts of COVID-19-related stay-at-home orders through a two-way random effects model. *Transp Res Rec*. 2023;2677(4):255–266. doi:10.1177/03611981211046921
19. Ten Hove D, Jorgensen TD, van der Ark LA. Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychol Methods*. 2024;29(5):967–979. doi:10.1037/met0000516
20. Li M, Gao Q, Yu T. Methodological issues on statistical rigor of agreement analysis. *Scientometrics*. 2022;128(3):2025–2027. doi:10.1007/s11192-022-04591-4
21. Yeates P, Maluf A, Kinston R, et al. Enhancing authenticity, diagnosticity and equivalence (AD-Equiv) in multicentre OSCE exams in health professionals education: protocol for a complex intervention study. *BMJ Open*. 2022;12(12):e064387. doi:10.1136/bmjopen-2022-064387
22. Peng Q, Luo J, Wang C, Chen L, Tan S. Impact of station number and duration time per station on the reliability of Objective Structured Clinical Examination (OSCE) scores: a systematic review and meta-analysis. *BMC Med Educ*. 2025;25(1):84. doi:10.1186/s12909-025-06691-0
23. Humphrey-Murto S, Shaw T, Touchie C, Pugh D, Cowley L, Wood TJ. Are raters influenced by prior information about a learner? A review of assimilation and contrast effects in assessment. *Adv Health Sci Educ Theory Pract*. 2021;26(3):1133–1156. doi:10.1007/s10459-021-10032-3
24. Andersen SAW, Nayahangan LJ, Park YS, Konge L. Use of generalizability theory for exploring reliability of and sources of variance in assessment of technical skills: a systematic review and meta-analysis. *Acad Med*. 2021;96(11):1609–1619. doi:10.1097/ACM.0000000000004150
25. Grüter AAJ, Van Lieshout AS, van Oostendorp SE, et al. Video-based tools for surgical quality assessment of technical skills in laparoscopic procedures: a systematic review. *Surg Endosc*. 2023;37(6):4279–4297. doi:10.1007/s00464-023-10076-z
26. Yeates P, McCray G, Moult A, Cope N, Fuller R, McKinley R. Determining the influence of different linking patterns on the stability of students' score adjustments produced using Video-based Examiner Score Comparison and Adjustment (VESCA). *BMC Med Educ*. 2022;22(1):41. doi:10.1186/s12909-022-03115-1
27. Fu Y, Zhang W, Zhang S, Hua D, Xu D, Huang H. Applying a video recording, video-based rating method in OSCEs. *Med Educ Online*. 2023;28(1):2187949. doi:10.1080/10872981.2023.2187949
28. Yeates P, Maluf A, Cope N, et al. Using video-based examiner score comparison and adjustment (VESCA) to compare the influence of examiners at different sites in a distributed objective structured clinical exam (OSCE). *BMC Med Educ*. 2023;23(1):803. doi:10.1186/s12909-023-04774-4
29. Ross SB, Modasi A, Christodoulou M, et al. New generation evaluations: video-based surgical assessments: a technology update. *Surg Endosc*. 2023;37(10):7401–7411. doi:10.1007/s00464-023-10311-7
30. Tan JY, Ma IWY, Hunt JA, et al. Video recording in veterinary medicine OSCEs: feasibility and inter-rater agreement between live performance examiners and video recording reviewing examiners. *J Vet Med Educ*. 2021;48(4):485–491.
31. Hara S, Ohta K, Aono D, et al. Feasibility and reliability of the pandemic-adapted online-onsite hybrid graduation OSCE in Japan. *Adv Health Sci Educ Theory Pract*. 2024;29(3):949–965. doi:10.1007/s10459-023-10290-3
32. Tavares W, Kinnear B, Schumacher DJ, Forte M. "Rater training" re-imagined for work-based assessment in medical education. *Adv Health Sci Educ Theory Pract*. 2023;28(5):1697–1709. doi:10.1007/s10459-023-10237-8
33. Makrides A, Yeates P. Memory, credibility and insight: how video-based feedback promotes deeper reflection and learning in objective structured clinical exams. *Med Teach*. 2022;44(6):664–671. doi:10.1080/0142159X.2021.2020232
34. Pattni C, Scaffidi M, Li J, et al. Video-based interventions to improve self-assessment accuracy among physicians: a systematic review. *PLoS One*. 2023;18(7):e0288474. doi:10.1371/journal.pone.0288474