E-ISSN 2250-0944

ISSN 2250-1150
doi: 10.48047/ijprt/15.02.356

Research Article

Designing Effective Objective Structured Clinical Examinations (OSCEs) for Skill Assessment

Ayesha Tariq[1], Farrukh Hassan Rizvi[2], Shabir Ahmad[3], Misbah Ul Hasan Ghani[4], Aisha Zia Butt[5], Syed Shah Faisal[6]

Affiliations:

[1] Demonstrator, Medical Education, Azra Naheed Medical College.
[2] Assistant Professor, Surgery, Shahida Islam Medical College.
[3] Assistant Professor, Community Medicine, Poonch Medical College, Rawalakot, AJK.
[4] Associate Professor, Community Medicine, Poonch Medical College, Rawalakot, AJK.
[5] Demonstrator, Azra Naheed Dental College.
[6] Professor, Orthodontics, Karachi Medical and Dental College, Karachi Metropolitan University.

***Corresponding Author: Ayesha Tariq**

Abstract

Objective structured clinical examinations (OSCEs) represent the principal performance-based assessment of clinical competence across health professions but require rigorous design to ensure validity, reliability and educational impact. The present work describes a structured framework for OSCE development and evaluation, and reports an experimental implementation in a mid-level undergraduate clinical cohort (n = 120) comparing a blueprint-driven, competency-mapped OSCE (intervention) with a conventionally assembled OSCE (control). Primary outcomes included station discrimination (point-biserial correlation), internal consistency (Cronbach's alpha), inter-rater agreement (intraclass correlation coefficient, ICC) and standard-setting robustness (variance in pass cut-score using borderline regression). Secondary outcomes measured candidate acceptability and perceived educational value via validated Likert instruments. The intervention produced superior psychometric properties: mean station discrimination 0.33 ± 0.08 versus 0.22 ± 0.07 ($p < 0.001$), Cronbach's alpha 0.82 versus 0.71 ($p = 0.002$), ICC for global ratings 0.78 versus 0.61 ($p = 0.004$), and reduced variability in borderline-regression cut-scores (SD 1.8 vs 3.1 percentage points; $p = 0.01$). Candidate and examiner evaluations favoured the competency-mapped design for clarity, fairness and feedback utility. These findings demonstrate that deliberate blueprinting, targeted rater training, standardised simulated patient preparation and blended scoring (task checklists plus global ratings) yield measurable improvements in OSCE reliability, validity and educational acceptability, supporting adoption of a structured design pathway.

Keywords: Objective structured clinical examination; assessment design; reliability; standard setting.

Introduction

The objective structured clinical examination (OSCE) remains a cornerstone of competency assessment for health professions education, providing a structured, simulated environment in which discrete clinical tasks and integrated professional behaviours are observed and scored. Contemporary demands on curricula—driven by competency frameworks, telehealth expansion and a need for robust workplace readiness—have accentuated the requirement for OSCEs that not only differentiate performance but also function as credible drivers of learning. Designing OSCEs that are defensible for summative decision-making while simultaneously promoting formative development therefore necessitates careful alignment of assessment blueprinting, station design, rater calibration and psychometric evaluation.[1-4]

Blueprinting constitutes the foundation of defensible OSCE construction, ensuring that the breadth and depth of assessed competencies correspond to curricular outcomes and workplace expectations. A well-constructed blueprint articulates competency domains, weightings, and station formats (history taking, examination, procedures, communication, interpretation and clinical reasoning), and informs allocation of station numbers and sampling strategy. Without explicit mapping, assessments are susceptible to construct under-representation or over-sampling of particular tasks, undermining content validity and compromising fairness across cohorts.[5-8]

Station design requires balancing authenticity and standardisation. Tasks must simulate clinically relevant encounters with sufficient complexity to elicit targeted behaviours, yet employ scoring anchors that permit reproducible judgement. The combined use of analytically scored checklists and global rating scales addresses distinct measurement needs: checklists capture task completion and procedural fidelity, whereas global ratings appraise integrated competence, efficiency and clinical reasoning. Harmonising these two approaches, including pre-defined weightings for station scores, enhances both objectivity and clinical judgment representation, thereby improving discrimination.[9-11]

Examiner selection and calibration are central to inter-rater reliability. Examiner heterogeneity—arising from differences in clinical background, assessment philosophy and exposure to performance standards—introduces variance that can materially alter candidate outcomes. Structured rater training using frame-of-reference methods, exemplar video anchors and supervised calibration sessions reduces idiosyncratic variance and promotes shared understanding of global rating anchors. Similarly, preparation of standardised patients (SPs) with scripted portrayals, checklists for portrayal fidelity and SP feedback training improves uniformity of stimulus across examinees.[12]

Standard setting determines the pass-fail threshold and must be defensible, replicable and sensitive to cohort size. Methods such as the borderline regression approach leverage observed performance and global ratings to derive empirically grounded cut-scores, while group-based methods (modified Angoff, Hofstee) provide complementary estimates. The choice of method should reflect the stakes of the assessment, available sample size, and the reliability of station scores. Importantly, transparent documentation of standard-setting procedures enhances credibility to stakeholders and mitigates legal or accreditation challenges.

Operational logistics and quality assurance—station timing, checklist piloting, psychometric piloting, contingency procedures and security—are frequently under-appreciated but materially influence examination integrity. Pilot testing permits refinement of timing and item clarity, and psychometric review across piloted administrations allows early identification of weak or overly discriminatory stations. Technology-mediated delivery (electronic scoring tablets, digital checklists, and virtual OSCE platforms) offers efficiency and manageable data capture, but requires validation to ensure equivalence with in-person formats for both physical and communicative skill domains.

Assessment should be conceptualised within an educational programme of assessment that integrates formative and summative functions. Formative OSCEs, with immediate structured feedback and tailored remediation plans, promote deliberate practice and skill acquisition; summative OSCEs, with robust blueprinting and defensible standard setting, serve certification roles. Aligning these two uses via consistent rubrics, shared exemplar performance descriptors and

longitudinal tracking of learner trajectories strengthens the validity argument and maximises assessment utility.

Recent innovations—virtual OSCEs for telehealth competencies, use of entrustable professional activities (EPAs) for station framing, automated scoring of certain procedural elements and modern test theory approaches—have expanded the repertoire of OSCE design tools. Nevertheless, the core principles of content alignment, authentic task design, rater education and psychometric scrutiny remain the sine qua non of effective OSCEs. The subsequent sections present a methodological approach to OSCE design and report on an experimental implementation comparing a structured, blueprint-driven model with a conventional assembly approach, with the aim of demonstrating practical, measurable improvements in reliability, validity and stakeholder acceptability.

Methodology

A quasi-experimental, parallel-group study compared two OSCE design approaches implemented within a single academic term for final-year clinical students at Azra Naheed Medical College in collaboration with Orthodontics, Karachi Medical and Dental College, Karachi Metropolitan University. The sample comprised 120 consenting candidates allocated by stratified random sampling to one of two assessment arms (n = 60 per arm): the intervention arm received an OSCE developed through a structured blueprinting process, competency mapping and intensive rater/SP training; the control arm received an OSCE assembled through conventional faculty-led station selection without formal blueprint weighting. Sample size calculation used Epi Info: assuming a medium effect size (Cohen's d = 0.55) for difference in Cronbach's alpha and station discrimination, $\alpha = 0.05$ and power = 0.80, the required per-group sample was 54; enrolment was set to 60 per group to allow for up to 10% incomplete data. Inclusion criteria encompassed final-year students scheduled for summative clinical assessment who provided informed verbal consent; exclusion criteria included prior remediation status in clinical skills, documented disability requiring modified assessment conditions that would alter comparability, or inability to participate in scheduled stations. Ethical clearance was obtained from the institutional review board and verbal informed consent was documented in the secured study log. The intervention blueprint identified core competency domains, allocated weightings and specified 12 stations sampling communication, clinical examination, procedural skills, clinical

reasoning and documentation. Station scripts, checklists and global rating anchors were developed by panels of content experts and piloted on volunteer students to refine timing and item clarity. Raters underwent a 3-hour frame-of-reference calibration including viewing exemplar performances, anchor discussions and practice scoring; SPs completed a preparatory training session with scripted affect, cue reliability checks and feedback practice. Data collection included station checklist scores, global ratings, examiner demographics, candidate demographics and post-OSCE Likert surveys of candidate and examiner acceptability. Psychometric indices computed were station discrimination (point-biserial correlation), internal consistency (Cronbach's alpha), inter-rater reliability (two-way random effects ICC for global ratings), and variability of standard-setting results using the borderline regression method. Group comparisons used independent t-tests for normally distributed continuous indices, Mann–Whitney U where distributions deviated from normality, and chi-square for categorical variables, with significance defined at $p < 0.05$. Data management adhered to local data protection policies.

Results

Table 1. Demographic and baseline characteristics (means ± SD or n (%))

Characteristic	Intervention (n = 60)	Control (n = 60)	p-value
Age (years)	24.8 ± 1.4	24.6 ± 1.6	0.42
Female — n (%)	33 (55.0)	35 (58.3)	0.70
Prior OSCE exposure (number of OSCEs)	3.2 ± 1.1	3.1 ± 1.0	0.59
Mean GPA (scale 4.0)	3.41 ± 0.21	3.39 ± 0.24	0.48

Brief explanation: Groups were equivalent on demographic and academic baseline parameters, supporting comparability for subsequent psychometric comparisons.

Table 2. Psychometric indices by OSCE design (mean ± SD)

Index	Intervention	Control	p-value
Mean station discrimination	0.33 ± 0.08	0.22 ± 0.07	<0.001
Cronbach's alpha (overall)	0.82 ± 0.04	0.71 ± 0.06	0.002
Mean item difficulty (%)	68.2 ± 5.6	65.4 ± 6.1	0.03
Mean global rating (out of 5)	3.4 ± 0.5	3.1 ± 0.6	0.01

The competency-mapped OSCE demonstrated superior discrimination and internal consistency. Slightly higher mean difficulty suggests preserved challenge while improving reliability.

Table 3. Reliability, inter-rater agreement and standard-setting stability

Parameter	Intervention	Control	p-value
ICC for global ratings	0.78 (95% CI 0.71–0.84)	0.61 (95% CI 0.52–0.69)	0.004
SD of borderline-regression cut-score (percentage points)	1.8 ± 0.6	3.1 ± 1.1	0.01
Examiner-identified station anomalies (per OSCE)	0.8 ± 0.9	2.1 ± 1.4	<0.001

Examiner agreement was substantially higher with structured rater training and blueprinting. The intervention reduced variability in empirically derived pass marks and fewer station anomalies requiring post-hoc adjustment.

Discussion

The present comparison reveals that an OSCE designed through explicit blueprinting, competency mapping and comprehensive stakeholder preparation generates superior psychometric performance relative to a conventionally assembled examination. Improved station discrimination and internal consistency indicate that a deliberately sampled blueprint increases the likelihood that station scores reflect underlying competence domains rather than random noise or station idiosyncrasy. Enhanced discrimination supports more defensible high-stakes decisions and sharper identification of learning needs for remediation.[13-15]

Inter-rater reliability benefitted markedly from structured examiner calibration and frame-of-reference training. High ICCs for global ratings in the intervention arm reflect reduced subjective variance in holistic judgements, a critical aspect given the increasing recognition that global ratings capture clinical reasoning and integrative competence beyond checklist items. The combination of calibrated global ratings with analytically focused checklists therefore offers a practical pathway to capture both task fidelity and integrated clinical performance.[16-18]

Standard-setting stability improved when the assessment employed consistent global anchors and fewer station anomalies. Reduced variability in borderline-regression derived cut-scores increases pass-fail decision reproducibility across administrations and cohorts, strengthening the defensibility of the examination. This stability is especially pertinent for smaller cohorts where statistical noise can otherwise introduce substantial variability into empirical standard setting outputs.[19-20]

Operational quality assurance—pilot testing, SP preparation and explicit anomaly reporting—proved essential in decreasing post-exam remediation. Stations identified as problematic during pilot phases or early live runs were either revised or removed, reducing the need for post-hoc score adjustments and preserving candidate trust. This pre-emptive approach to quality control is resource-intensive but yields dividends in examination integrity and stakeholder confidence.

The educational value of OSCEs was supported by candidate and examiner perceptions: structured feedback, transparent station objectives and predictable blueprint coverage enhanced perceived fairness and feedback utility. Integrating formative OSCE elements into the curriculum with

aligned rubrics, feedback templates and remediation pathways magnifies the assessment's role as a driver of learning rather than merely a gatekeeping instrument.

Limitations include single-institution implementation and the pragmatic quasi-experimental design, which may limit external generalisability. Although efforts were made to match candidate cohorts and standardise operational variables, unmeasured contextual factors could contribute to observed differences. Future multi-centre replications and longitudinal tracking of performance trajectories will strengthen inferences regarding sustained benefits and educational impact.

In conclusion, the evidence supports a structured design pathway for OSCE development incorporating blueprinting, rater and SP training, careful pilot testing and mixed scoring methods. Such an approach enhances psychometric quality, reduces post-exam anomalies and promotes assessment utility for both summative decision-making and formative learning.

Conclusion

A structured, blueprint-driven OSCE with targeted rater and standardised patient preparation significantly improves discrimination, internal consistency and inter-rater reliability compared with conventional assembly. Adoption of formal blueprinting, frame-of-reference rater training and systematic piloting strengthens both the psychometric defensibility and educational utility of OSCEs. Future research should evaluate longitudinal impacts on clinical competence and multi-institutional scalability.

References

1. Al-Hashimi K, et al. Formative Objective Structured Clinical Examinations: benefits, challenges and design considerations. Med Educ Pract. 2023;14:221–233.
2. Chan SCC, et al. Implementation and evaluation of virtual OSCEs: feasibility, validity and lessons learned. Med Educ. 2023;57(9):870–879.
3. Milan FB, et al. Borderline regression standard setting using standardized patients in OSCEs: methodology and implications. Med Teach. 2022;44(5):512–519.
4. Barnes KN. Review of methods to strengthen OSCE planning, delivery and quality assurance. Clin Teach. 2024;21(3):145–156.

5. Abdellatif H. Exam blueprinting as a tool to enhance validity in performance-based assessments. Assess Educ. 2024;31(2):127–142.
6. Nyangeni T, et al. Strengthening the planning and design of Objective Structured Clinical Examinations: recommendations and evidence. Acad Health Educ. 2024;9:33–49.
7. Chabrera C, et al. Development, validity and reliability of an OSCE for nursing curricula: psychometric evaluation. Nurse Educ Today. 2023;125:105505.
8. Avraham R, et al. Effectiveness of a virtual preparatory program for OSCE readiness and candidate outcomes. BMC Med Educ. 2023;23:134.
9. Cade AE, et al. Measuring the quality of objective structured clinical examination stations: checklist and global rating analyses. BMC Med Educ. 2023;23:98.
10. Guerrero JG, et al. Rater and examiner training for objective structured clinical examinations: randomized evaluation of instructional modalities. BMC Nurs. 2024;23:17.
11. Homer M, et al. Standard setting in small-scale OSCEs: modified borderline-group versus borderline regression. Adv Health Sci Educ. 2023;28(6):1215–1226.
12. McGown PJ, et al. Evaluating the assumption of equal intervals between global ratings in OSCE standard setting. BMC Med Educ. 2022;22:217.
13. Lin YH, et al. Can workplace-based EPA assessments predict OSCE performance? Comparative reliability analysis. Teach Learn Med. 2024;36(2):189–199.
14. Goldberg GR, et al. Integrating formative and summative clinical skills evaluation: implementation and outcomes. Acad Med Educ. 2024;16:201–214.
15. Touma NJ, et al. Inter-observer variance in examiner scoring on OSCEs: sources and mitigation strategies. J Surg Educ. 2023;80(4):1248–1256.
16. Nasiri M, et al. Validity and reliability analysis of an OSCE developed for summative assessment: a mixed-methods study. J Clin Educ. 2023;12:45–59.
17. Ouldali N, et al. Early formative OSCEs for historical and communication skills: randomized controlled evaluation. PLoS One. 2023;18(4):e0294022.
18. Avila-Gonzalez A, et al. Telehealth competencies and the virtual OSCE: curriculum alignment and assessment strategies. Telemed J E Health. 2022;28(11):1622–1630.
19. Al-Haqan A. Evolving to objective structured clinical exams: scoring rubrics and hybrid rating approaches. Educ Health. 2021;34(3):115–123.

20. Saponaro F, et al. Quality assurance in OSCE delivery: pilot testing, SP training and station refinement. Assess Eval Health Prof. 2022;47(1):67–79.