

Enhancing the Reliability and Validity of the Objective Structured Clinical Examination (OSCE): A Multi-Faceted Intervention Study at Kufa University Faculty of Medicine

Ahmed Athab Al Zubaidy¹, Alaa Jumaah Manji Nasrawi^{*2}, Waseem Ahmed El Katib³, Habib Shahib Al-Hassani⁴, Rajaa Rashid Abbas⁵, and Fadhaa A. Ghafil⁶

^{1,3,4} Department of Surgery, University of Kufa, Faculty of Medicine, Iraq.

² Department of Pediatrics, University of Kufa, Faculty of Medicine, Iraq.

⁵ Department of Physiology, University of Kufa, Faculty of Medicine, Iraq.

⁶ Department of Pharmacology, University of Kufa, Faculty of Medicine, Iraq.

*Corresponding Author Email: ahmedam.alzubaidi@uokufa.edu.iq, * alaaj.nasrawi@uokufa.edu.iq, waseem.alkatib@uokufa.edu.iq, habeeb.almohammed@uokufa.edu.iq, rajaa.alsaffar@uokufa.edu.iq, fadhaa.alhadrawi@uokufa.edu.iq

ABSTRACT

Background: The Objective Structured Clinical Examination (OSCE) is a key component of the test battery used to evaluate performance in simulated settings. This research evaluates the reliability and validity of the OSCE as a component of the Final Professional Exam (FPE) at the University of Kufa, Faculty of Medicine, for the academic year 2023-2024. **Aims:** The study aimed to enhance OSCE reliability by refining station design, examiner training, and the examination environment, in accordance with the standards set by the International Association for Health Professions Education (AMEE). **Methods:** A cross-sectional study was conducted at the University of Kufa, Faculty of Medicine, between January and June 2024. Four faculty workshops on OSCE design and assessment were delivered in accordance with AMEE guidelines. The OSCE was conducted at 24 stations, each lasting 5 minutes, across four parallel circuits, with 28 trained simulated patients. Spearman's rank correlation and Cronbach's Alpha were used to assess validity and reliability. **Results:** The OSCE had 210 students, and 98.571% passed. It showed high reliability (Cronbach's Alpha = 0.95). There was a moderate positive correlation between the OSCE and written exam scores ($r = 0.605$). Compared with the previous year's exam (IPE 2), reliability and correlation were higher, indicating that the interventions were effective. However, the between-group variance (32.8%) and the weak correlation between sets 1 and 4 (-0.134) indicate areas that need more work. **Conclusion:** The study emphasizes the necessity of stringent validation in educational assessments to maintain confidence in examination outcomes and advocates for ongoing efforts to improve OSCE reliability and student readiness, especially in communication skills.

Keywords: OSCE, Reliability, Validity, Cronbach's Alpha, Correlation.

Article Information

Received: December 06, 2025; Revised: May 04, 2026; Online June, 2026

INTRODUCTION

Epstein (2007) posits that a combination of assessment tools can evaluate various learning domains; this methodology is referred to as a 'TEST BATTERY' approach. The Objective Structured Clinical Examination (OSCE) is a crucial component of this test battery for assessing

performance in simulated environments ⁽³⁾. The OSCE was first created in the 1970s because there was a need for a more realistic and objective way to test clinical skills than traditional tests could offer ⁽⁴⁾. Before the implementation of OSCEs, evaluations during clinical rotations frequently relied on subjective assessments influenced by

factors such as patient availability and the designated preceptor for each student.

The following crucial components determine OSCEs' success: (1) careful **station design** is vital to ensure that scenarios authentically represent real-world clinical settings and the abilities being tested; (2) it is essential to have uniform **scoring rubrics** with well-defined criteria to ensure uniformity in examiner judgments; (3) a group of **skilled examiners**; and (4) creating a realistic and captivating examination atmosphere depends heavily on trained **simulated patients** ⁽³⁾.

OSCE was adopted in 2013 at the Kufa Faculty of Medicine, and multiple post-exam psychometric analyses since then have pointed towards poor reliability and low correlation with the written exam. The medical education unit hypothesized that improving OSCE performance could strengthen the correlation and reliability of the written exam. The primary aim was to enhance the reliability and validity of the Final Professional Exam (FPE) OSCE by improving station design, examiner training, and the exam environment, in line with AMEE standards.

METHODS

This cross-sectional study was conducted at the University of Kufa, Faculty of Medicine, Iraq, between January and June 2024. The Kufa faculty adopted an integrated curriculum in 2013, consisting of two phases: Phase One, which lasts six semesters and includes integrated, clinically applied basic sciences; and Phase Two, which lasts three semesters and comprises pure clinical subjects with horizontal integration. At the end of the sixth stage, there is a high-stakes Final Professional Exam (FPE) consisting of a written exam and an OSCE.

The medical education unit adopted a plan to improve the FPE OSCE reliability by working on four points (intervention): station design, examiner training, improving the exam environment to resemble real life, and well-

trained simulated patients. Four workshops were conducted by experts in medical education and assessment from the medical education department at the University of Kufa, Faculty of Medicine. These workshops were directed to all faculty teachers involved in OSCE in groups over three months. They covered a historical and theoretical perspective on OSCE, its definition, variants, validity, reliability, feasibility, educational impact, comparison with other assessment tools, blueprinting, station types, station rubric, piloting, global rating, briefing for both students and assessors, and psychometric analysis. The AMEE guidelines for OSCE guided all these workshops ⁽³⁾.

In May, the OSCE exam was held across 24 stations, each lasting 5 minutes. One teacher rated each station. The OSCE was conducted in four sets, with students randomly assigned to them. It was designed using a blueprint that covered multiple topics and clinical skills (see Table 1). After suitable training, 28 real and simulated patients participated. The OSCE stations were written according to a standardized template. The global rating rubric was: poor, borderline, pass, very good, and excellent.

We hypothesized that we could achieve good reliability, a strong correlation between the theory exam and the OSCE, and improvements in these parameters relative to the previous summative exam for the same student group at the 5th stage of 2022-2023, namely Intermediate Professional Exam 2 (IPE 2).

The data were collected and analyzed using Microsoft Excel 2019. Spearman's rank correlation was used to assess the associations between the checklist and global rate scores, and between the written exam and the OSCE. Before analysis, data were screened for normality. Because the OSCE scores showed a non-normal distribution—common in high-stakes clinical assessments—non-parametric statistics were employed. Spearman's rank correlation was selected as it is appropriate for ordinal data and

does not require the assumption of normality. Cronbach's Alpha was calculated to assess internal consistency. The interpretation of the correlation coefficient is shown in Table 2.

The OSCE ran in four sets (circuits) simultaneously. The students were divided into four groups in alphabetical order. In each group, about 53 students completed the same checklist with a different examiner; a single examiner conducted each station.

Table 1: The FPE OSCE 2024 Blueprint.

Station	Topic	Hx	CE	DI	Comm	Proc	Rad	Pharm
1	Job Shadow (Sx)							☞
2	Job Shadow (Mx)		☞					
3	Job Shadow (Px)	☞						
4	Job Shadow (GO)				☞			
5	Acute Care 1		☞					
6	Acute Care 2					☞		
7	Chronic Care 1			☞				
8	Chronic Care 2							☞
9	Psychiatry	☞						
10	Neurology		☞					
11	Oncology 1		☞					
12	Oncology 2				☞			
13	GIT			☞				
14	MSK		☞					
15	Vascular		☞					
16	Plastic					☞		
17	Urology/NeuroS _x				☞			
18	Cardiac		☞					
19	Respiratory	☞						
20	Metabolic				☞			
21	Pediatric 1						☞	

Station	Topic	Hx	CE	DI	Comm	Proc	Rad	Pharm
22	Pediatrics 2					☞		
23	Gyne/Obst 1	☞						
24	Gyne/Obst 2							☞
T	24 stations	4	7	4	4	3	2	2

Hx = History; CE = Clinical Examination; DI = Data Interpretation; Comm = Communication skills; Proc = Procedural skills; Rad = Radiology; Pharm = Pharmacology.

Table 2: Interpretation of correlation coefficient ⁽⁵⁾.

Size of correlation	Interpretation
0.9 to 1 (-0.9 to 1)	Very high positive (negative)
0.7 to 0.9 (-0.7 to -0.9)	High positive (negative)
0.5 to 0.7 (-0.5 to -0.7)	Moderate positive (negative)
0.3 to 0.5 (-0.3 to -0.5)	Low positive (negative)
0 to 0.3 (-0 to -0.3)	Negligible correlation

Cronbach's Alpha was used to measure the reliability of the exam. Table 3 shows the interpretation of Cronbach's Alpha.

Table 3: The interpretation of Cronbach's Alpha ⁽⁶⁾.

Cronbach's Alpha	Interpretation
$\alpha > 0.9$	Excellent
$\alpha > 0.8$	Good
$\alpha > 0.7$	Acceptable
$\alpha > 0.6$	Questionable
$\alpha > 0.5$	Poor
$\alpha < 0.5$	Unacceptable

Ethical Considerations

All participants provided written informed consent before their involvement in the study. The study protocol was approved by the Ethics Committee of Kufa University Faculty of Medicine (Protocol Number: #324), December 13, 2023. Participation was voluntary, and data confidentiality was maintained throughout the research.

RESULTS

In the Final Professional Exam (FPE, 2023-2024), 210 students took the OSCE, and 207 passed (98.571%). The pass mark was adjusted according to the standard setting using borderline regression; see Table 4.

Table 4: Descriptive statistics for each OSCE station.

Station	Mean	SD	Max	Min	Cut score	Pass %
Psychiatry	7.40	1.23	10	4.5	4	100
MSK	7.01	1.37	10	2	5.5	91.4
Metabolic	7.46	1.80	10	1	5	89.5
GIT	7.0	1.74	10	2.5	6	75.2
Cardio respiratory	6.45	1.54	10	3	5	81.9
Cardio respiratory	7.89	1.12	10	5	5	100
Plastic Surgeon	6.61	1.83	10	2	6	66.1
Urology Neurosurgery	7.21	2.10	10	0.5	6	73.8
Vascular	7.47	1.71	10	2.5	6	82.3
Neurology	8.25	1.55	10	2.5	5	96.6
Rheumatology	7.84	1.87	10	1	6	87.1
Radiology	8.67	1.19	10	2	5.8	98.5
Communication	6.26	1.57	9	1	5	84.7
Acute Care	6.790	1.4	10	2.5	6	81.4
Acute Care	7.176	1.89	10	2.5	2.5	100
Nephrology	5.498	2.74	10	0.5	5	52.8
Cardiac	7.629	1.67	10	2.5	5	93.8
Hematology	8.119	1.32	10	4.5	7	82.8
Pediatric	6.483	1.78	10	1.5	6	66.6
Pediatric	7.531	2.20	10	0	6.3	81.9
Pediatric	5.94	1.65	9.474	0	4.8	75.7
Gynecology	7.721	1.70	10	2	5	93.8
Gynecology	6.613	1.99	10	1	6.5	61.4
Obstetric	7.933	1.81	10	0	6.5	84.7

The OSCE exam was comprehensive, covering several disciplines and skills. The success rates across the four central departments are shown in Figure 1.

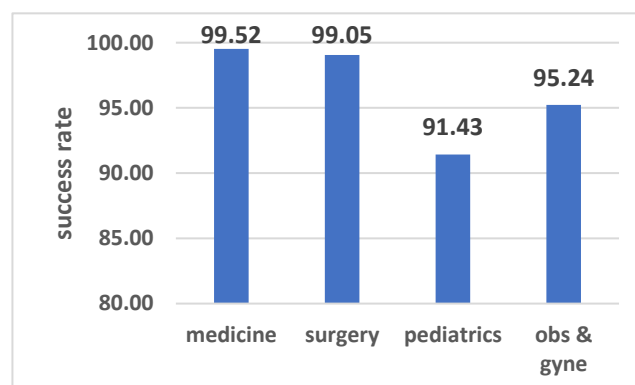


Figure 1: The success rate in the four main clinical departments.

According to the tested skills, the success rate is shown in Figure 2.

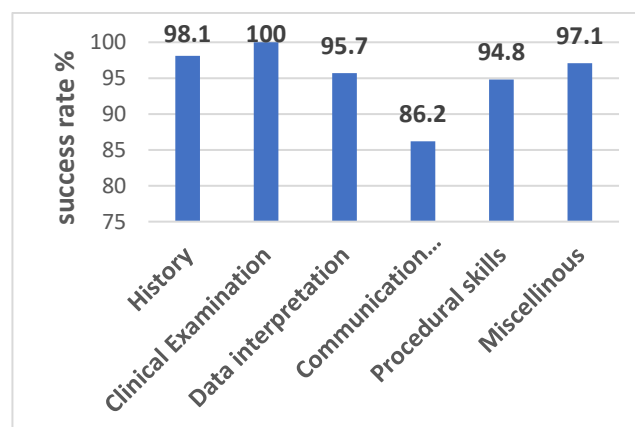


Figure 2: The success rate of the six main clinical skills tested in OSCE.

The success rates for each station and each set are shown in Table 5.

Table 5: The success rate according to each station and set.

Topic	Set 1	Set 2	Set 3	Set 4
Psychiatry	100	100	100	100
MSK	76	90.7	100	98.1
Metabolic	98	92.6	88.7	79.2
GIT	92	81.5	81.1	45.3
Cardiorespiratory	90	94.4	75.5	67.9
Cardiorespiratory	100	100	100	100
Plastic surgery	90	66.7	54.7	54.7

Topic	Set 1	Set 2	Set 3	Set 4
Urology / Neurosurgery	92	63	62.3	79.2
Vascular	100	87	90.6	52.8
Neurology	98	90.7	98.1	100
Rheumatology	98	83.3	81.1	86.8
Radiology	98	100	96.2	100
Communication	96	88.7	84.9	69.8
Acute Care	66	84.9	86.8	86.8
Acute Care	100	100	100	100
Nephrology	100	43.4	43.4	26.4
Cardiac	82	98.1	98.1	96.2
Hematology / Oncology	64	88.7	98.1	79.2
Pediatric radiology	82	39.6	73.6	71.7
Pediatric resuscitation	84	83	64.2	96.2
Pediatric job shadow	100	77.4	62.3	64.2
Gynecology	92	94.3	98.1	90.6
Gynecology	36	69.8	52.8	84.9
Obstetric	88	79.2	77.4	94.3
Average	88.4	83.2	82	80.2

The correlation among the four sets (four examiners for the same station) was studied in Table 6.

Table 6: Correlation among the four OSCE sets.

	Set 1	Set 2	Set 3	Set 4
Set 1	1			
Set 2	0.6137	1		
Set 3	0.6082	0.7901	1	
Set 4	-0.1347	0.1929	0.1792	1

- **Strongest correlation:** between Set 2 and Set 3 (0.79), indicating a strong positive relationship.
- **Weakest correlation:** between Set 1 and Set 4 (-0.134), indicating a fragile negative relationship.

- **Inter-Class Correlation Coefficient:** 0.378.
- **Between-group variance:** 32.8%.

The correlation (R^2) between the global rating and the checklist score was examined in Figure 3.

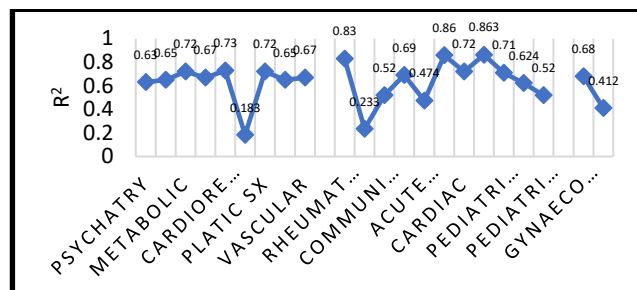


Figure 3: The correlation between global rating and checklist score for each OSCE station.

The weakest correlations were observed at the cardiorespiratory and radiology stations, with values of 0.183 and 0.233, respectively. For the neurology and gynecology stations, no R^2 values were calculated because the examiners did not sign the global rating.

Cronbach's alpha for the OSCE exam was 0.95, indicating high reliability.

The Ministry of Higher Education held the written exam, supervised by the committee of medical faculty deans. It was a computer-based exam using Mizan Company's (UR) assessment platform in 2024. Students had to choose from surgical specialties, medical specialties, gynecology and obstetrics, and pediatrics in 100 single-best-answer items per paper, with a one-week interval between papers. As students could compensate across these four exams, the cumulative success rate in this exam was 97.62% (205 out of 210). The passing percentage per paper was highest in Obstetrics and Gynecology and lowest in Pediatrics; see Figure 4.

Figure 4: The success rate of each main department in the unified ministerial theory exam.

Cronbach's alpha for this written exam was 0.935, indicating high reliability. The correlation

between this written exam and the OSCE was 0.605 (Figure 5).

Figure 5: The correlation between the written exam and OSCE in FPE, 2023-2024.

In the Intermediate Professional Exam 2 (IPE 2), the summative exam for the same student group at the end of the 5th stage 2022-2023, 189 out of 206 students passed the written exam (91.95%) compared with 100% for the OSCE. The correlation between the two exams was 0.476 (Figure 6). Cronbach's Alpha for OSCE was 0.678.

Figure 6: The correlation between the written exam and OSCE in IPE 2, 2022-2023.

Core Findings of the OSCE Evaluation

- Significant improvement in reliability.
- Strong correlation with theory.
- High student success rate.
- Skill-specific performance variation.

DISCUSSION

The study showed that the OSCE held at the FPE in the University of Kufa, Faculty of Medicine, during 2023-2024, was highly reliable (Cronbach's Alpha = 0.95). Compared with the OSCE held in 2022-2023 for the same student cohort (IPE 2; Cronbach's Alpha = 0.678), this represents a significant improvement, primarily attributable to the medical education unit's efforts to enhance the reliability of the FPE OSCE. Moreover, the results showed a stronger correlation between theoretical and OSCE scores in the FPE ($r = 0.605$) than in IPE 2 ($r = 0.476$).

Rigorous validation of educational assessments is essential because those who use them must trust the results. Early pioneers of OSCE claim that high content validity is achievable in this format, especially when using a 'test blueprint' ⁽⁴⁾. According to Moattari et al., the criterion-referenced validity of the OSCE is confirmed by the positive and significant correlation between OSCE and theoretical exam

scores ⁽⁷⁾. Gerrow et al. examined the criterion validity of the OSCE by evaluating dental practical skills using written board exam scores from 2,317 senior students over five years; they showed a positive and significant relationship between written exam scores and OSCE results ^(8,9). Sabzi et al. found a higher correlation between theoretical and OSCE scores than between clinical and OSCE scores ⁽¹⁰⁾.

For logistical reasons, students are separated into groups for the OSCE. This separation has to be entirely random. In a perfect evaluation process, all grade variations would result from variations in student performance rather than from variations in location, environment, or assessor attitudes (e.g., 'hawks and doves') ⁽¹¹⁾. The between-group variation should ideally be less than 30%; values higher than 40% should raise red flags. In this study, the Inter-Class Correlation Coefficient was 0.378, and the variance was 32.8% between groups, with poor correlation between sets 1 and 4 (-0.134). All these statistics raise concern and require further investigation by reviewing assessor performance and considering recalibration. Homer ⁽¹²⁾ describes a higher correlation of 0.75 among OSCE circuit groups, although he compared only two circuits among 442 examinees in an 18-station OSCE for international medical graduates. Yeates et al. ⁽¹³⁾ report a score variance of up to 11%, much lower than in this study; however, they studied the same OSCE with different examinee groups, using 2011 UK Foundational Program data and the 2008 University of Alberta Multiple Mini-Interview.

The R^2 coefficient represents the proportion of the change in the dependent variable (checklist score) explained by the independent variable (global grade). A high correlation ($R^2 > 0.5$) indicates a strong association between checklist scores and global grades ⁽¹¹⁾. In Figure 3, the cardiorespiratory and radiology stations show poor correlations (0.183 and 0.233, respectively), indicating that some students received many marks on the item checklist. At the same time,

their overall performance raised concerns for the assessor, leading to a failing grade. This prompts a review of the item checklist and station design.

Internal structure validity evidence involves analyzing the OSCE's different psychometric properties⁽¹⁴⁾, as shown in Table 1 and Figures 1, 2, 3, and 4. The assessment committee must review these statistics and take action. The lowest success rate among the six primary clinical skills tested was for communication skills, with higher rates in clinical examinations (86.2% and 100%, respectively). Sim et al.⁽¹⁵⁾ studied OSCE student performance across history-taking, physical examination, communication, clinical reasoning, procedural skills, and professionalism, with the lowest score in clinical reasoning. Following the well-known phrase 'assessment drives learning'⁽¹⁶⁾, we recommend improving student training in communication skills at the Kufa Faculty of Medicine.

CONCLUSION

The combined efforts of the medical education unit—faculty workshops, revised blueprinting, and use of trained simulated patients—led to a clear improvement in OSCE reliability at the Kufa Faculty of Medicine. Cronbach's Alpha rose from 0.678 in IPE 2 (2022–2023) to 0.95 in FPE (2023–2024), and the correlation with written exam scores improved from 0.476 to 0.605. Because station redesign, examiner training, and environmental changes were introduced together, we cannot determine how much each intervention contributed.

Study Limitations

- **Confounding variables:** Because all four interventions were introduced simultaneously, we cannot tell which one had the greatest effect on the improvement in reliability.
- **Inter-circuit variability:** The between-group variance was 32.8%, and the correlation between sets 1 and 4 was poor

(-0.134). These figures suggest that the 'hawks and doves' effect among examiners persists despite the training workshops.

- **Data gaps:** Examiners at the neurology and gynecology stations did not sign the global rating, so R^2 could not be calculated for those stations.

Specific Recommendations

- **Communication-skills training:** Communication skills had the lowest success rate (86.2%) among the six tested skills. We recommend adding more dedicated communication training during Phase Two of the clinical curriculum.
- **Assessor recalibration:** The R^2 between checklist scores and global ratings was only 0.183 (cardiorespiratory) and 0.233 (radiology). These stations require examiner recalibration so that checklist marks better match the examiner's overall impression of student competence.
- **Consistency monitoring:** With an ICC of 0.378 and between-group variance of 32.8%, we recommend monitoring examiner scoring in real time during the OSCE so that circuit-based discrepancies can be caught and corrected on the day.
- **Longitudinal tracking:** It would be valuable to follow this cohort through their remaining assessments to see whether the gains in reliability and correlation persist over time.

ACKNOWLEDGMENT

The authors thank the Medical Education Unit of the University of Kufa, Faculty of Medicine, for facilitating the OSCE workshops, and all faculty teachers, simulated patients, and students who participated in the Final Professional Exam 2023-2024.

Ethical approval

The present study was approved by the Ethics Committee of Kufa University Faculty of Medicine (Protocol Number: #324) on December 13, 2023.

Statement of Permission and Conflict of Interests

The authors declare no conflicts of interest related to this study.

REFERENCES

1. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387–96.
2. Hamdy H, Telmesani AW, Wardy NA, Abdel-Khalek N, Carruthers G, Hassan F, et al. Undergraduate medical education in the Gulf Cooperation Council: a multi-country study (Part 2). *Med Teach*. 2010;32(4):290–5.
3. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: a historical and theoretical perspective. *Med Teach*. 2013;35(9):e1437–46.
4. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using an objective structured examination. *BMJ*. 1975;1(5955):447–51.
5. Ahammad P. How to interpret correlation coefficients [Internet]. Parvez Ahammad's Website; 2016 [cited 2024 Jun 4]. Available from: <http://www.parvez-ahammad.org/blog/how-to-interpret-correlation-coefficients>
6. George D, Mallery P. SPSS for Windows step by step: a simple guide and reference, 11.0 update. 4th ed. Boston: Allyn and Bacon; 2003.
7. Moattari M, Abdollah-Zargar S, Mousavinasab M, Zare N, Marvdast PB. Reliability and validity of OSCE in evaluating clinical skills of nursing students. *Iran J Med Educ*. 2009;13(3).
8. Fallatah HI, Tekian A, Park YS, Shawa LA. The validity and reliability of the sixth-year internal medical examination administered at the King Abdulaziz University Medical College. *BMC Med Educ*. 2015;15(1):190.
9. Gerrow JD, Murphy HJ, Boyd MA, Scott DA. Concurrent validity of written and OSCE components of the Canadian dental certification examinations. *J Dent Educ*. 2003;67(8):896–901.
10. Sabzi Z, Modanloo M, Yazdi K, Kolagari S, Aryaie M. The validity and reliability of the objective structured clinical examination (OSCE) in pre-internship nursing students. *J Res Dev Nurs Midwifery*. 2018;15(1):1–9.
11. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: a review of metrics—AMEE Guide No. 49. *Med Teach*. 2010;32(10):802–11.
12. Homer M. Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes. *Adv Health Sci Educ Theory Pract*. 2022;27(2):457–73.
13. Yeates P, Moreau M, Eva K. Do contrast effects influence examiners' judgments in OSCE-style assessments? *Acad Med*. 2015;90(7):975–80.
14. Hodges B. Validity and the OSCE. *Med Teach*. 2003;25(3):250–4.
15. Sim JH, Aziz YFA, Mansor A, Vijayananthan A, Foong CC, Vadivelu J. Students' performance in the different clinical skills assessed in OSCE: what does it reveal? *Med Educ Online*. 2015; 20:26185.
16. Ludvigsson JF. Ramsden on learning to teach in higher education. *J Pediatr Gastroenterol Nutr*. 2003;36(4):511–2.