Title

Optimizing cost-effectiveness in remote objective structured clinical examinations through targeted double scoring methodologies.

Permalink
https://escholarship.org/uc/item/8d81j4n5

Journal
Medical Education Online, 30(1)

Authors
Fu, Zhihui
Wu, Yuhong
Xu, Lingling
et al.

Publication Date
2025-12-01

DOI
10.1080/10872981.2025.2467477

Peer reviewed

MEDICAL EDUCATION ONLINE
2025, VOL. 30, 2467477
https://doi.org/10.1080/10872981.2025.2467477

RAPID COMMUNICATION

OPEN ACCESS

Optimizing cost-effectiveness in remote objective structured clinical examinations through targeted double scoring methodologies

Zhihui Fu [a,b], Yuhong Wu [a,b], Lingling Xu [c,d], Fen Cai [c,d], Ren Liu [e] and Zhehan Jiang [c,d]

[a]Department of Statistics, School of Mathematics and Statistics, Minnan Normal University, Zhangzhou, Fujian, China; [b]Fujian Key Laboratory of Granular Computing and Applications, Minnan Normal University, Zhangzhou, Fujian, China; [c]Institute of Medical Education, Peking University, Beijing, China; [d]Peking University Health Science Center-Chaoxing Joint Laboratory for Digital and Smart Medical Education, Beijing, China; [e]Psychological Science, University of California, Merced, CA, USA

ABSTRACT

The remote Objective Structured Clinical Examination (OSCE) is a cornerstone of medical education, enabling structured and objective assessment of clinical skills, communication, and patient-centered care. However, its widespread adoption has introduced challenges related to cost-effectiveness and efficient use of rater resources. Traditional double scoring (DS) ensures reliability but is labor-intensive and costly, especially in large-scale assessments. To address these challenges, this study introduces Targeted Double Scoring (TDS), a novel methodology that selectively applies DS to specific score ranges, particularly those near the pass/fail threshold. The study was conducted using data from a pilot remote OSCE administered to 550 clinical medicine undergraduates in China. The OSCE consisted of three stations: Clinical Reasoning (CR), Physical Examination (PE), and Fundamental Skills (FS). Each station was scored remotely by two raters, with a cut-off score of 60 out of 100. The TDS methodology was modeled based on the OSCE's DS design and fitted with scoring data. A decision-theoretic approach identified optimal Critical Score Ranges (CSRs) for targeted double scoring, balancing reliability and cost-effectiveness. The findings show that TDS significantly reduces rater workload and costs while maintaining high reliability and fairness. For instance, TDS achieved up to 70% cost savings compared to traditional DS under certain configurations. The study also highlights the flexibility of TDS, which can be tailored to different OSCE designs and scoring rubrics. These results have broad implications for medical education, especially in resource-constrained settings where optimizing assessment efficiency is critical. This study provides a practical solution to the cost-related challenges of remote OSCEs and offers a framework for adopting TDS in assessments. By focusing raters on critical score ranges, TDS maintains rigorous and fair evaluations without overburdening faculty or exceeding budgets. Future research should explore TDS scalability and its integration with emerging technologies like artificial intelligence to enhance efficiency and reliability.

ARTICLE HISTORY
Received 7 December 2023
Revised 21 January 2025
Accepted 11 February 2025

KEYWORDS
OSCE; remote scoring; targeted double scoring; cost-saving strategy; digital assessment

Introduction

The importance of performance assessment in medical education cannot be overstated. Performance assessment, such as the Objective Structured Clinical Examination (OSCE), is a valuable tool for medical educators to assess the knowledge, skills, and attitudes of medical students in a clinically relevant context. OSCEs, particularly those conducted remotely, have become a cornerstone of medical education, enabling the assessment of clinical skills, communication, and patient-centered care in a structured and objective manner [1]. OSCEs provide a structured, objective method of assessing the performance of medical students that can be used to give feedback on their progress. Additionally, OSCEs can be used to assess the effectiveness of medical education programs, provide guidance for curriculum changes, and identify areas for further improvement [2].

Traditional (onsite) scoring of OSCEs has been the standard for many years; it involves having trained faculty and/or standardized patients assess students' performance in real-time. Typically, raters have evaluated students in close proximity – by their side, in the same room, or through a one-way mirror. Despite the advantages of remote scoring, traditional double scoring (DS) can be costly and time-consuming, particularly for large-scale assessments. This study addresses the gap by introducing Targeted Double Scoring (TDS), which selectively applies DS to critical score ranges to optimize cost-effectiveness [1]. Further, social facilitation theory implies that the presence of raters can lead to downward performance when tasks are complex [3,4]. Last but not least, the onsite costs of onsite scoring can be substantial, resulting in a heavy budgetary load in most well designed OSCEs [5]; recent literature has addressed issues for cost saving in OSCE scoring [6,7]. Largely

CONTACT Zhehan Jiang jiangzhehan@gmail.com; Lingling Xu xllpsy@qq.com Institute of Medical Education, Peking University, Beijing, China; Peking University Health Science Center-Chaoxing Joint Laboratory for Digital and Smart Medical Education, Beijing, China.

reducing the aforementioned burdens, remote scoring has become increasingly popular in the medical education field.

Remote scoring involves using digital tools to evaluate students' performances, either for OSCEs conducted in person or fully remote OSCEs, using predetermined criteria [1]. Remote scoring has the potential to trim the costs and improve the efficiency of performance assessment while still providing accurate feedback to students. There has been a growing body of research on the use of remote scoring in medical education. For example, a recent study by Wu et al. [8] explored the use of remote scoring for an OSCE assessment in a medical education program. The study found that remote scoring was reliable and valid for assessing the performance of medical students, and that it had the potential to reduce costs and improve the efficiency of the assessment process. Similarly, a study by Arnold et al. [9] evaluated the use of remote scoring for a clinical skills assessment. The study found that remote scoring had good reliability and validity, and that it was a cost-effective and time-efficient method of assessing medical students' performance.

Moreover, several studies highlight the challenges and benefits associated with implementing OSCEs. Majumder et al. [10] noted that although OSCEs are widely recognized as effective assessment tools, they require significant resources for administration and evaluation. Chong et al. [11] emphasized the importance of examiner training and standardization to ensure consistency in scoring. Kim et al. [12] developed and evaluated an OSCE designed to assess medical students' clinical performance, underscoring the need for clear criteria and robust methods to maintain assessment quality. Brand and Schoonheim-Klein [13] reported that different assessment methods could influence examination anxiety among students, suggesting that careful consideration should be given to the psychological impact of various assessment formats. Alfaris et al. [14] examined student perceptions of OSCEs at a new medical school, revealing that despite initial apprehension, students generally viewed OSCEs positively once familiarized with the format. Park et al. [15] compared examinee perceptions between internal and external examiners, indicating that external examiners might offer more objective evaluations. Lastly, Gopalan et al. [16] provided insights into how students perceive OSCEs, offering recommendations for enhancing the assessment experience.

It should be noted that, definition-wise, remote scoring is a general concept, within which a scoring assignment (i.e., who and how to rate students) strategy plays a critical role in driving the workflow. Among different scoring assignment strategies, double scoring (DS) involves having two different raters independently score the same set of responses, and then comparing the two scores to determine the final score [17]. This method is especially useful when dealing with non-structured responses, as it (1) helps to reduce the potential for bias in scoring and (2) allows for an independent comparison of the two results. The advantages of DS are more evident when the scoring rubric is complex or when the rubric does not provide clear-cut guidelines for the scoring [16]. However, scoring of performance tasks tends to be expensive and time-consuming [18]. To control the costs, DS might be performed for a random subset of students instead of all. Finkelman et al. [19] suggested that a better approach would be targeted DS (TDS), which emphasizes DS of those students who are close to the pass/fail point and can be accurately identified; this approach is particularly useful when the target assessment involves standard settings (i.e., a cut-off is demanded) [1].

An important term that one should understand is the Critical Score Range (CSR), referring to a specific range of scores where the risk of incorrect pass/fail decisions is highest. For example, if the pass/fail cutoff is 60, scores close to 60 (e.g., 55–65) are considered critical because small differences in these scores can significantly impact a student's outcome. By focusing double scoring on these critical ranges, we ensure fairness and reliability without unnecessary cost. One can regard the CSR as a 'red zone' in a game. Just as players need extra caution when entering the red zone, scores near the pass/fail cutoff require extra scrutiny. By focusing double scoring on this 'red zone,' we ensure that critical decisions are made with high reliability.

Figure 1 presents a streamlined visualization of the TDS concept using a simplified example. In this example, the total score ranges from 0 to 20, with a pass/fail cutoff at 8. The Critical Score Range (CSR) is highlighted in red, showing that scores close to 8 (e.g., 6–10) require double scoring to ensure fairness. Scores outside this range (e.g., 0–5 or 15–20) do not require double scoring, as the risk of incorrect decisions is lower. Under DS, examinees P1 and P2 would both require assessments from two raters. However, with TDS implementation, only select examinees would need double rating, deemed necessary based on a single rater's sufficiently reliable score, thus eliminating the need for universal DS application. This targeted approach significantly reduces the overall rating effort and associated costs. A critical aspect of TDS involves identifying the CSR that indicates examinees near the pass/fail threshold. Accordingly, scores within a CSR warrant additional raters to ensure assessment fairness and reliability. As demonstrated, different CSRs imply varying rating requirements; for example, while P2 falls within CSR.1 necessitating two raters, P1 does not. Conversely,

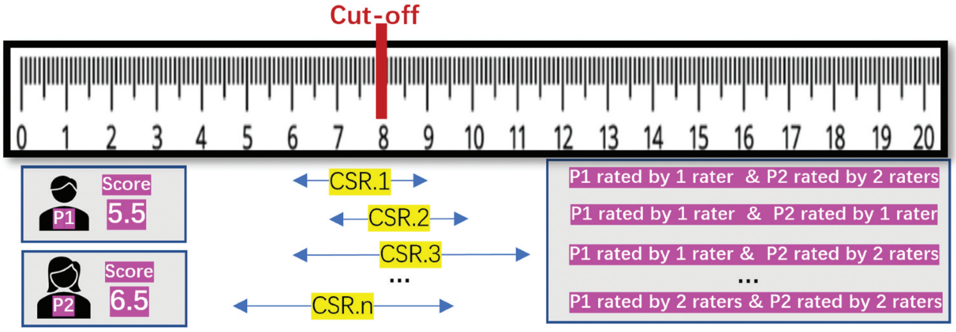

Cut-off
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Score
P1
5.5
Score
P2
6.5
CSR.1
CSR.2
CSR.3
...
CSR.n
P1 rated by 1 rater & P2 rated by 2 raters
P1 rated by 1 rater & P2 rated by 1 rater
P1 rated by 1 rater & P2 rated by 2 raters
...
P1 rated by 2 raters & P2 rated by 2 raters

Figure 1. Demonstration of targeted double scoring in a simplified scenario when different critical score ranges (Csr.X) are present.

neither falls within CSR.2, indicating both require only one rater. Finally, since both scores are within CSR.n, double rating is applied to ensure thorough evaluation for both examinees. Note that the case shown in Figure 1 is indeed a simplified one, as in practice assessment tends to contain multiple tasks and/or stations such as OSCEs. TDS taking place at the task- or station-level, instead of overall score, would complicate the decision-making process because much more combinations (e.g., DS is required for the first and the second tasks but unneeded for the third task) for each examinee emerge in the analysis.

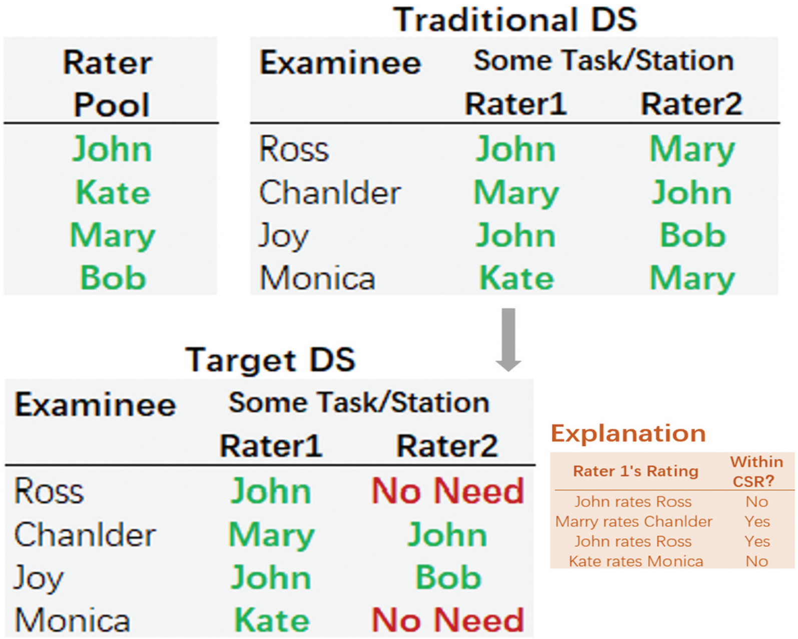
In this paper, we adopt the Targeted Double Scoring (TDS) methodology, which builds on the principles of decision-theoretic approaches to assessment [20] and addresses the limitations of traditional DS. In the OSCE's setting, as each station contains only one task, the station and the task are treated interchangeable throughout the analysis. Therefore, it's appropriate the state that each station/task contains multiple scoring points by raters in this study. Specifically, the TDS is modeled according to the OSCE's DS design and is fitted with the scoring data. Armed with quantitative decision-making process initially proposed by Sinharay and colleagues [20], TDS is expected to optimize the selection of examinees whose responses need to be scored by two raters. Therefore, the TDS results are meant to provide a guideline as well as a cost-effective solution for generalizing the remote scoring practice to OSCEs using the similar setting.

From an intervention standpoint, applying TDS in OSCE assessments can be considered a 'treatment', given that the original OSCE format involves two or more raters. This approach has a dual objective: firstly, to evaluate whether TDS offers a more cost-effective alternative to the current scoring setup; and secondly, if the initial hypothesis holds true, to determine which CSR yields the most optimal outcomes. Therefore, this paper serves as a methodological paper with a typical OSCE setting to help readers realize the benefits of TDS and the feasibility of adopting TDS to their own OSCEs.

Methods

In a typical OSCE involving fail/pass decisions, two or more raters are required to maintain higher reliability and consistency, as rater variability can significantly impact scoring outcomes. TDS begins with one rater, labelled as Rater 1, who is randomly selected from the rater pool with a specific label. to ensure unbiased initial scoring. That said, those labels identify the expertise of each rater, ensuring that the raters only rate performance that they have superior professional qualities. That said, each examinee's responses to a task are scored by a rater, termed as Rater 1 whoever completes the initial rating assignment. Therefore, Rater 1 is likely to be unidentical for different examinees in an actual OSCE, because multiple yet parallel testing rooms are always available in the setting, while two examinees starting the test simultaneously are scored by different Rater 1s. Figure 2 demonstrates how four raters, selected for the rater pool, assess the performance of four examinees on specific tasks. Regardless of the DS method utilized, Examinee Ross is initially evaluated by Rater John, followed by a subsequent assessment from Rater Mary, making John Ross's initial rater (i.e., Rater 1), a role he shares with Examinee Joy. For Examinees Chandler and Monica, their initial raters are Mary and Kate, respectively. The figure highlights that Ross and Monica's performances, as assessed by Rater 1s (John and Kate), fall outside the CSR, negating the need for a second evaluation. To emphasize, the Rater (e.g., Rater 1 or Rater 2) is a label in the TDS methodological workflow; this conceptual term is used in the statistical formula of TDS loss function as seen later in this paper.

According to the rule demonstrated in Figure 1, TDS requires those examinees, of whom a score assigned by Rater 1 falling within a CSR, to be double-scored at the specific task/item. To distinguish the terms and concepts in this paper, Figure 3 illustrates



Traditional DS

Rater Pool
John
Kate
Mary
Bob

Examinee	Some Task/Station	
	Rater1	Rater2
Ross	John	Mary
Chanlder	Mary	John
Joy	John	Bob
Monica	Kate	Mary

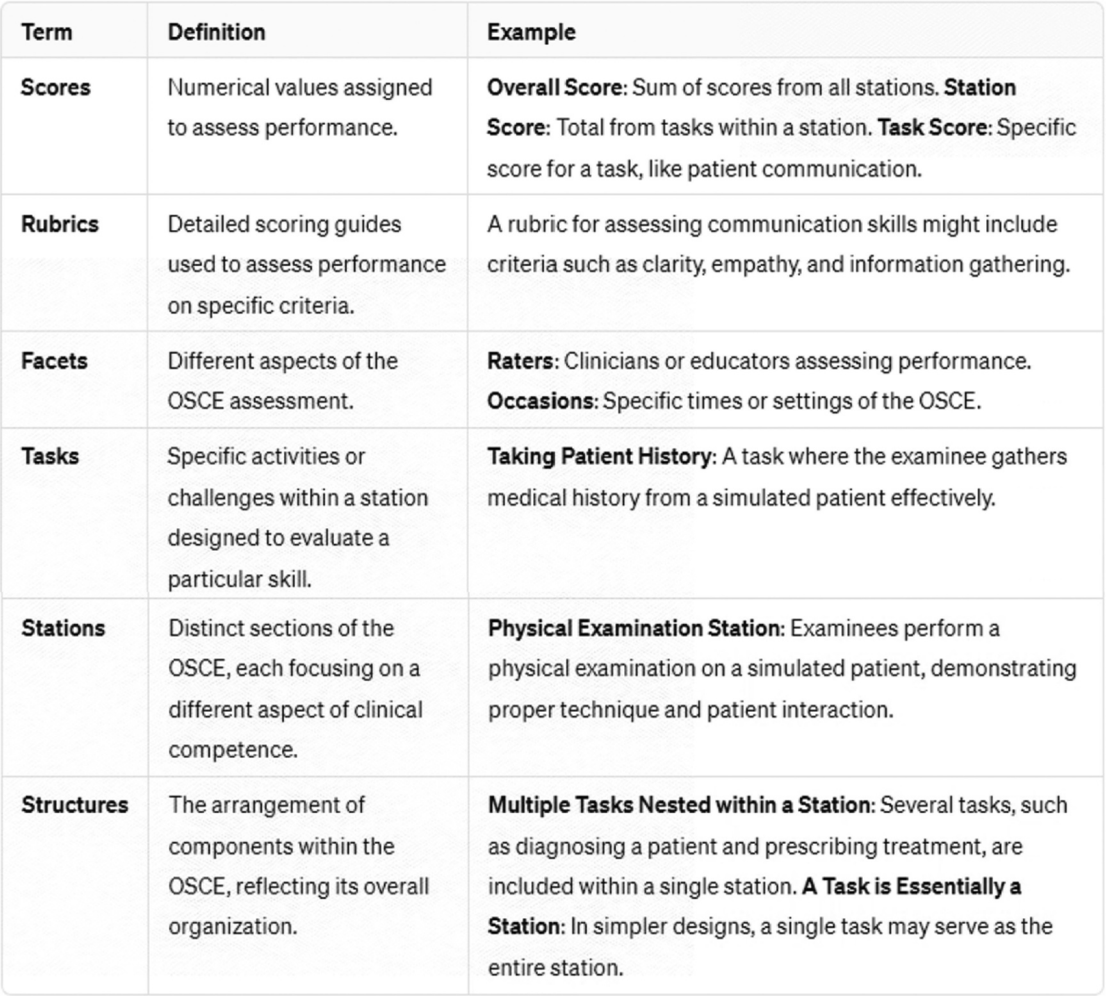
Target DS

Examinee	Some Task/Station	
	Rater1	Rater2
Ross	John	No Need
Chanlder	Mary	John
Joy	John	Bob
Monica	Kate	No Need

Explanation

Rater 1's Rating	Within CSR?
John rates Ross	No
Marry rates Chanlder	Yes
John rates Ross	Yes
Kate rates Monica	No

Figure 2. Illustration of the concept of Rater.



Term	Definition	Example
Scores	Numerical values assigned to assess performance.	**Overall Score**: Sum of scores from all stations. **Station Score**: Total from tasks within a station. **Task Score**: Specific score for a task, like patient communication.
Rubrics	Detailed scoring guides used to assess performance on specific criteria.	A rubric for assessing communication skills might include criteria such as clarity, empathy, and information gathering.
Facets	Different aspects of the OSCE assessment.	**Raters**: Clinicians or educators assessing performance. **Occasions**: Specific times or settings of the OSCE.
Tasks	Specific activities or challenges within a station designed to evaluate a particular skill.	**Taking Patient History**: A task where the examinee gathers medical history from a simulated patient effectively.
Stations	Distinct sections of the OSCE, each focusing on a different aspect of clinical competence.	**Physical Examination Station**: Examinees perform a physical examination on a simulated patient, demonstrating proper technique and patient interaction.
Structures	The arrangement of components within the OSCE, reflecting its overall organization.	**Multiple Tasks Nested within a Station**: Several tasks, such as diagnosing a patient and prescribing treatment, are included within a single station. **A Task is Essentially a Station**: In simpler designs, a single task may serve as the entire station.

Figure 3. Reference of terms related to settings and examples in this paper.

the relationships between key terms, their settings, and examples. The CSR is determined based on statistical analysis of the scoring data. Scores falling within the CSR are those closest to the pass/fail cutoff, where the risk of incorrect decisions is highest. For example, if the cutoff is 60, scores between 55 and 65 are considered critical. By focusing double scoring on these ranges, we ensure reliability while minimizing costs. The CSR is the cornerstone of TDS. By identifying and focusing on scores near the pass/fail cutoff, we ensure that the most critical decisions are made with high reliability. This approach reduces costs while maintaining fairness and accuracy. Should CSR apply to performance at station- or test-level becomes a typical question; it does not only depend on the OSCE's design but also the scoring rubric. In terms of the design, OSCEs involve various *facets* (e.g., raters, stations, tasks/items and occasions) and *structures* (e.g., multiple tasks nested within a station or a task is essentially a station). Scoring practice, such as the use of measure instruments, the calculation of overall scores, the granularity of examinees' performance being rated, and the scale of the scoring sheet, affects the CSR decision on the other hand.

The data used in this study were collected from a pilot remote OSCE administered to 550 clinical medicine undergraduates in China, with a diverse range of clinical expertise and educational backgrounds. The OSCE comprised three stations: Clinical Reasoning (CR), Physical Examination (PE), and Fundamental Skills (FS). According to public information by NMLE, CR mainly focuses on the auscultation of the heart and lungs, imaging diagnosis, electrocardiogram diagnosis, medical ethics, medical history taking, and case analysis; among them, medical history taking and case analysis are paper-based, while the others are primarily computer-based. PE evaluates examinees' abilities to conduct comprehensive physical examinations on patients, covering various aspects such as general examination, head and neck examination, and chest examination; examinees are required to have a thorough understanding of examination methods, techniques, and judgment criteria. FS is designed to assess candidates' hands-on abilities, involving common clinical procedures such as injections, punctures, and intubations. Examinees need to proficiently master the essentials and precautions of these operations to ensure accurate and safe completion in practical situation.

Subscores were recorded for each station in the administration system, serving as the CSR levels for subsequent analysis. Each performance of the 550 examinees in the sample pool was rated remotely by two assessors. The cut-off score for the OSCE was set at 60 out of 100, which corresponds to the maximum possible total score. It should be noted that the OSCE did not record-specific cut-off subscores for each station or task. This study, involving human participants, received approval from the Biomedical Ethics Committee of Peking University (IRB00001052–22070).

In this paper, we assume that a CSR applies to tasks/stations of an OSCE instead of an overall score. Note that the scores on the remaining part of the OSCE may or may not be available at the time of deciding double score for one task/station. In addition, it is assumed that an overall OSCE score is obtained from combining scores on all tasks/stations; perhaps the simplest example is adding scores on all stations to form one value, which is compared with the cut-off score for pass/fail decisions. Finally, the unit of a score is equal to the average of two raters' judgements if double-scored examinees are present, else equal to the Rater 1 score.

In TDS, quantitative approaches [14–16] are used to aid defining CSR. A communality among the approaches is a loss function, which quantifies the benefit or loss associated with each possible decision. Traditionally, loss functions are defined to simple calculations of the monetary costs or the rater labor independently. Extending the loss function into the framework of statistical decision theory, on the other hand, allows the analysis to entail unknown factors, typically expressed as random variables and/or probability distribution(s) in a model. Compared with other quantitative counterparts, these model-based ones are likely to consider intake complex yet necessary elements into the consideration, yielding a more realistically useful and methodologically flexible solution. Therefore, loss functions underlaying statistical decision theory are used in this study. Sinharay and colleagues [20] consider three loss elements that are provided by researcher (i.e., the elements are not estimates but known constants used in the latter statistical process):

- L_P (if an examinee passes the OSCE based on Rater 1 score and would have failed the test if the subtest were double scored);
- L_F (if the examinee fails the OSCE based on Rater 1 score and would have passed if double-scored);
- c (the cost for extra rating when double scoring is needed).

Using $P()$ to represent a probability function and utilizing the abbreviations defined above, one can deliver the complete loss function as:

$$\begin{aligned}\text{Loss} = {} & c * P(\text{CSR}) + L_P \\ & * P(\text{Pass with 1rating and fail with2ratings } \& \text{Outside CSR}) \\ & + \ L_F \\ & * P(\text{Fail with 1rating and pass with2ratings } \& \text{Outside CSR}).\end{aligned}$$

$P(\text{CSR})$ presents the probability that a score falls in a selected CSR, for example, if 20% of examinees score 50–60 in an OSCE of which the score range is

0–100, then $P(\text{CSR}=[50,60]) = 0.2$; the remaining $P()$ components can be understood the same way. As the researcher provides values for L_P, L_F, and c, the goal of this study is to identify the CSR that results in the smallest loss, which can be considered the optimal CSR. It is important to clarify that TDS analysis is conducted on a complete dataset extracted from a past assessment. However, its implications pertain to a future design resembling a 'planned missing' approach. In this study, the value of c remained constant at 40, representing an average rating cost of $40 per performance. Given that L_P, and L_F cannot be directly obtained from the administration's financial statements, conventional combinations, as outlined by Sinharay and colleagues [20], were utilized to set the values as {40, 40}, {800, 200}, {400, 400}, {200, 800}, and {4,000, 4,000} for the costs of $\{L_P,L_F\}$. From a modeling perspective, the loss function defined above is the statistical model with both unknown estimates (i.e., CSR) and known constants (i.e., L_P,L_F, and c), and the computation is a combing plug-in and result comparison (i.e., trial and error algorithm). That is, different CSRs are suggested by experts and then used in the loss function to yield the loss estimates, which are then compared with each other to find the optimal solution.

In our exploration of cost savings and the loss function associated with the implementation of TDS, we aim to translate our statistical findings into more tangible, monetary terms for easier understanding. Essentially, the 'cost savings' we discuss refer to the reduction in expenses achieved by adopting TDS over traditional DS. For instance, if traditionally scoring an OSCE with two raters across all students costs $10,000, and implementing TDS reduces the need for dual raters to only critical cases, this might lower the cost to $6,000. This $4,000 difference represents our 'cost savings.' Similarly, the 'loss function' is a way of measuring what, if any, accuracy or reliability we sacrifice for these savings. However, our analysis aims to show that any such 'loss' is minimal and does not significantly impact the overall quality of the assessment. By presenting these concepts in terms of actual dollars saved, we hope to provide a clearer picture of TDS's economic and practical benefits.

Analysis

The full dataset is visualized in Figure 4, where each examinee' scores assigned by Rater 1, Rater 2, as well as the gap between the Raters (i.e., the lines connecting Rater 1 and Rater 2) are all shown. For points where no red lines are found, they mean that Rater 1 and Rater 2 reach a perfect agreement. Overall, the total scores rated by Rater 1 and Rater 2 form a Pearson correlation of 0.9 showing high total-score-level consistency, while computed via all raw scores, Cronbach's α reaches 0.66, an acceptable value for performance assessment. Using the aforementioned cut-off score (i.e., 60), employing Rater 1s' values alone results in a pass rate of 71.45%, while incorporating Rater 2s' data reduces the success rate to 68.72%. This reduction highlights the importance of double scoring in ensuring accurate pass/fail decisions, particularly in high-stakes assessments. The rating agreement between Rater 1 and Rater 2 was calculated using weighted Cohen's Kappa, yielding a value of 0.97, indicating near-perfect agreement. This statistic is considered high, with values above 0.80 generally indicating excellent agreement. Table 1 presents the descriptive findings of the remote OSCE. Notably, the patterns observed between Rater 1 and Rater 2 are quite similar, with the primary distinction between the pass and fail groups residing in the Clinical Reasoning (CR) station. From an alternative perspective, Rater 2's values exhibit a downward shift in the Patient Encounter (PE) and Formulation of Solutions (FS) stations when compared to those of Rater 1.

It is important to acknowledge that the values presented in Table 1 may undergo alterations when different combinations of raters are assigned to either Rater 1 or Rater 2 groups. Illustrating the fundamental outcomes of TDS, Figure 5 has been revised to include a more comprehensive table summarizing the expected losses for different CSRs, alongside the existing figure for better clarity. The CSR was divided into intervals based on the risk of incorrect decisions. For example, in the CR station, CSRs were divided into intervals of 10 points (e.g., 50–60); in the PE and FS stations, CSRs were divided into smaller intervals of 2 points (e.g., 58–60). These intervals were chosen to ensure that scores near the cutoff (60) were prioritized for double scoring. Each station encompasses two extremes of CSRs: the null set, signifying the absence of double scoring, and the set denoted as 0–60 for CR, 0–20 for PE, and 0–20 for FS, representing double scoring for all responses.

In Figure 5, each line corresponds to a specific CSR, with 'NULL' indicating the absence of double scoring throughout the assessment. The x-axis represents the estimated expected loss of the five sets of $\{L_P,L_F\}$, while the y-axis signifies the corresponding total cost, calculated using the Loss function as outlined in the Method section.

In cases where double scoring was not implemented (i.e., CSRs were set to 0–60 for CR, 0–20 for PE, and 0–20 for FS stations), the estimated expected loss consistently remained at 40, reflecting the value of 'c.' When $\{L_P,L_F\}$ equaled {40, 40}, the cost of incorrect pass/fail decisions did not surpass the cost of double scoring, rendering such incorrect decisions cost-effective. Consequently, the estimated expected loss

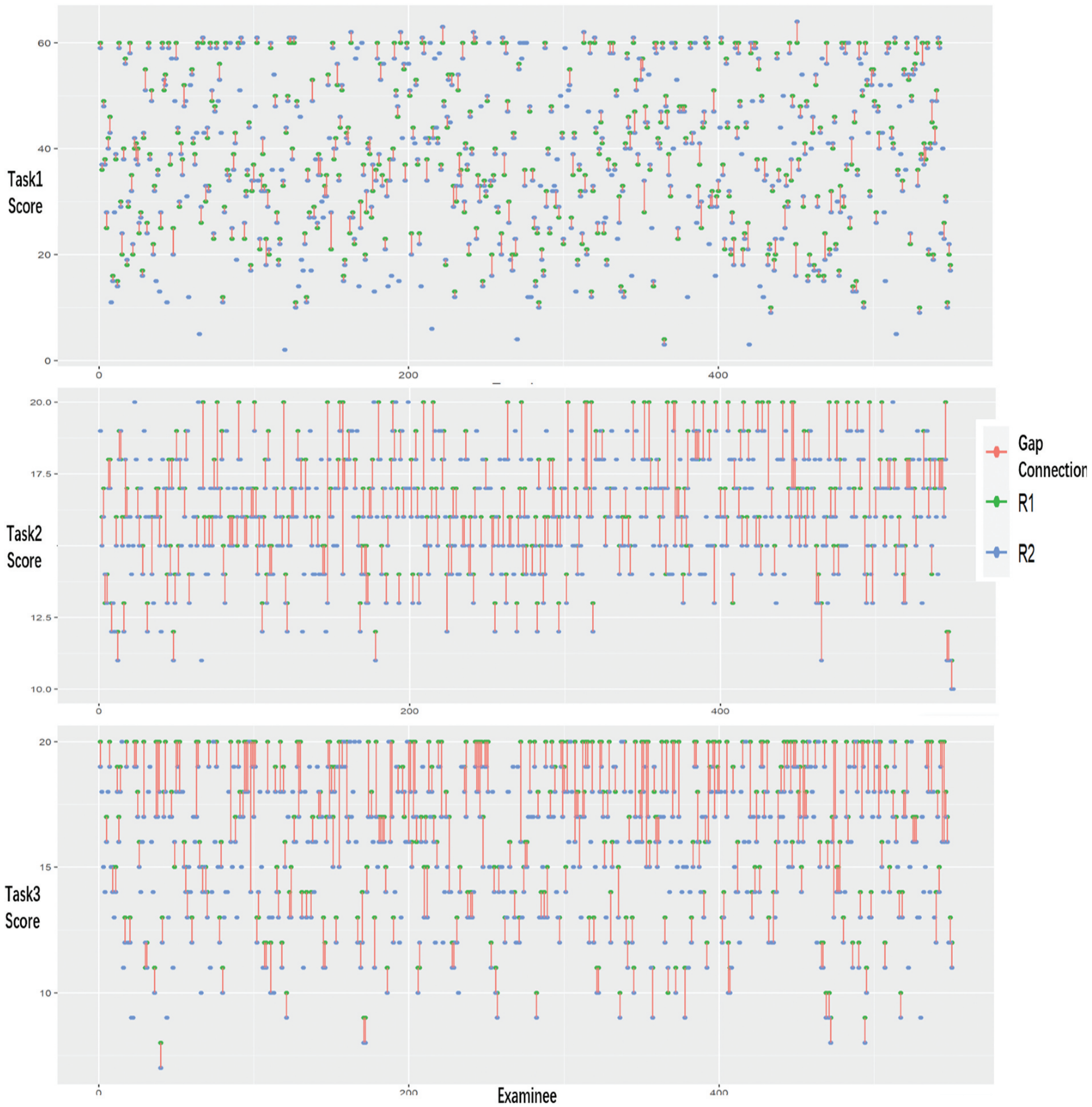


Figure 4. Plot of scoring results of each examinee by Raters.

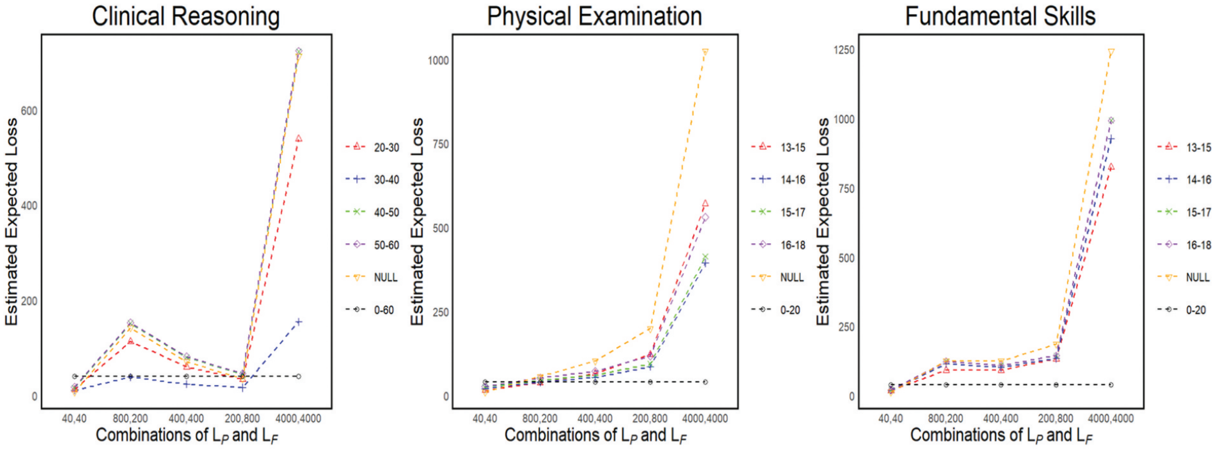
Table 1. Means and standard deviations of the remote OSCE.

		Test	Station		
		Total Score	Clinical Reasoning	Physical Examination	Fundamental Skills
Rater 1	Fail	49.70 (6.85)	19.24 (6.58)	15.24 (1.93)	15.22 (3.39)
	Pass	79.57 (11.06)	45.74 (10.41)	17.08 (1.78)	16.75 (2.92)
Rater 2	Fail	49.12 (7.32)	19.87 (7.08)	14.76 (1.95)	14.49 (3.18)
	Pass	77.95 (10.16)	45.68 (9.89)	16.47 (1.64)	15.81 (2.61)

was minimized for the null CSR, establishing ‘no double-scoring’ as the optimal decision for each station.

When L_P and L_F were set at 4,000, the loss incurred from incorrect pass/fail decisions exceeded the cost of Double Scoring (DS), rendering such incorrect decisions nearly unacceptable. Consequently, the estimated expected loss reached its minimum value (equating to 40) for the following CSRs: 0–20 in PE and FS stations, and 0–60 in CR station. This outcome indicates that opting for ‘all double-scoring’ was the optimal decision under these conditions.

For other sets of $\{L_P,L_F\}$, the estimated expected loss was minimized for the following CSRs: 30–40 in the CR station, 13–15 or 0–20 in the PE station, and 0–20 in the FS station. As the values of $\{L_P,L_F\}$ increased, the cost and risk associated with incorrect decisions rose, surpassing the cost of additional ratings. Consequently, the decision-theoretic approach favored DS and broader CSRs. This preference



- **X-axis:** Represents the estimated expected loss for different combinations of LP*LP* and LF*LF*. These combinations reflect varying costs associated with incorrect pass/fail decisions. For example, the x-axis includes values such as {40, 40}, {800, 200}, {400, 400}, {200, 800}, and {4,000, 4,000}.
- **Y-axis:** Represents the total cost, calculated using the loss function described in the Methods section. This cost includes both the monetary cost of double scoring (denoted as *c*) and the potential loss due to incorrect pass/fail decisions.
- **Lines in the Figure:** Each line corresponds to a specific CSR configuration for a given station. For example:
  - The line for "NULL" (no double scoring) represents the scenario where no scores are double-scored, regardless of their proximity to the pass/fail threshold.
  - Other lines represent different CSRs, such as 0-20, 30-40, etc., for each station. These CSRs indicate the score ranges where double scoring is applied to ensure reliability and fairness.

Figure 5. Estimated expected losses for three stations for different CSRs.

resulted in different CSR selections across the three combinations where L_P and L_F were equal. Specifically, the null range minimized the loss when $\{L_P,L_F\}$ equaled 40, while the full range was optimal when the unit cost in each station was set at 4,000. In the case of $\{L_P,L_F\}$ equaling 400, the middle range minimized the loss in the CR station, whereas the full range resulted in the smallest loss for the PE and FS stations.

Setting $\{L_P,L_F\}$ values to {40, 40}, {800, 200}, {400, 400}, {200, 800}, and {4,000, 4,000} correspondingly yielded minimum costs as follows: (1) in CR station, minimum costs were 7.13, 38.69, 24.15, 16.87, and 40; (2) in PE station, minimum costs were 10.25, 38.47, 40, 40, and 40; (3) in FS station, minimum costs were 12.44, 40, 40, 40, and 40. Summing the total minimum costs across all stations resulted in 29.82, 117.16, 104.15, 96.87, and 120, respectively. The {4,000, 4,000} combination preferred traditional DS over TDS, as the costs associated with the latter method were much higher, as seen from Figure 5.

Discussion and conclusion

Figure 6 summarizes the flow with the analysis results. (1) Initial Scoring: Rater 1 evaluated the examinee's performance and assigns a score. (2) Check if the score falls within the CSR: A decision point is reached to determine whether the score falls within the CSR. (3) If the score is outside the CSR, no double scoring is needed, else, Rater 2 was invited to evaluate the examinee's performance. (4) If DS was applied, the scores from Rater 1 and Rater 2 were combined (e.g., averaged) to determine the final result, else the Rater 1 score was used directly as the final result.

In exploring the implementation of TDS within the context of OSCEs, our study underscores the delicate balance between maintaining assessment integrity and managing logistical constraints. TDS emerges not only as a methodological innovation but as a strategic response to the increasing demands for scalable, cost-effective educational assessments. This approach, by focusing scoring resources on pivotal junctures within the assessment spectrum, particularly those close to the pass/fail threshold, ensures that the reliability of critical evaluations is upheld without disproportionately escalating administrative burdens. The implications of our findings extend beyond the immediate operational efficiencies. They invite a broader discourse on the evolution of assessment practices within clinical education, particularly in environments constrained by resources or facing logistical challenges. By demonstrating the feasibility and effectiveness of TDS, we contribute to a growing body of evidence advocating for adaptive assessment strategies. Such strategies, which harness data-driven insights to refine scoring methodologies, can significantly enhance the pedagogical value of OSCEs. They do so by ensuring that evaluations are not only rigorous and fair but also aligned with the practical realities of administering wide-scale examinations. Future research should explore the integration of TDS across diverse educational settings and its potential

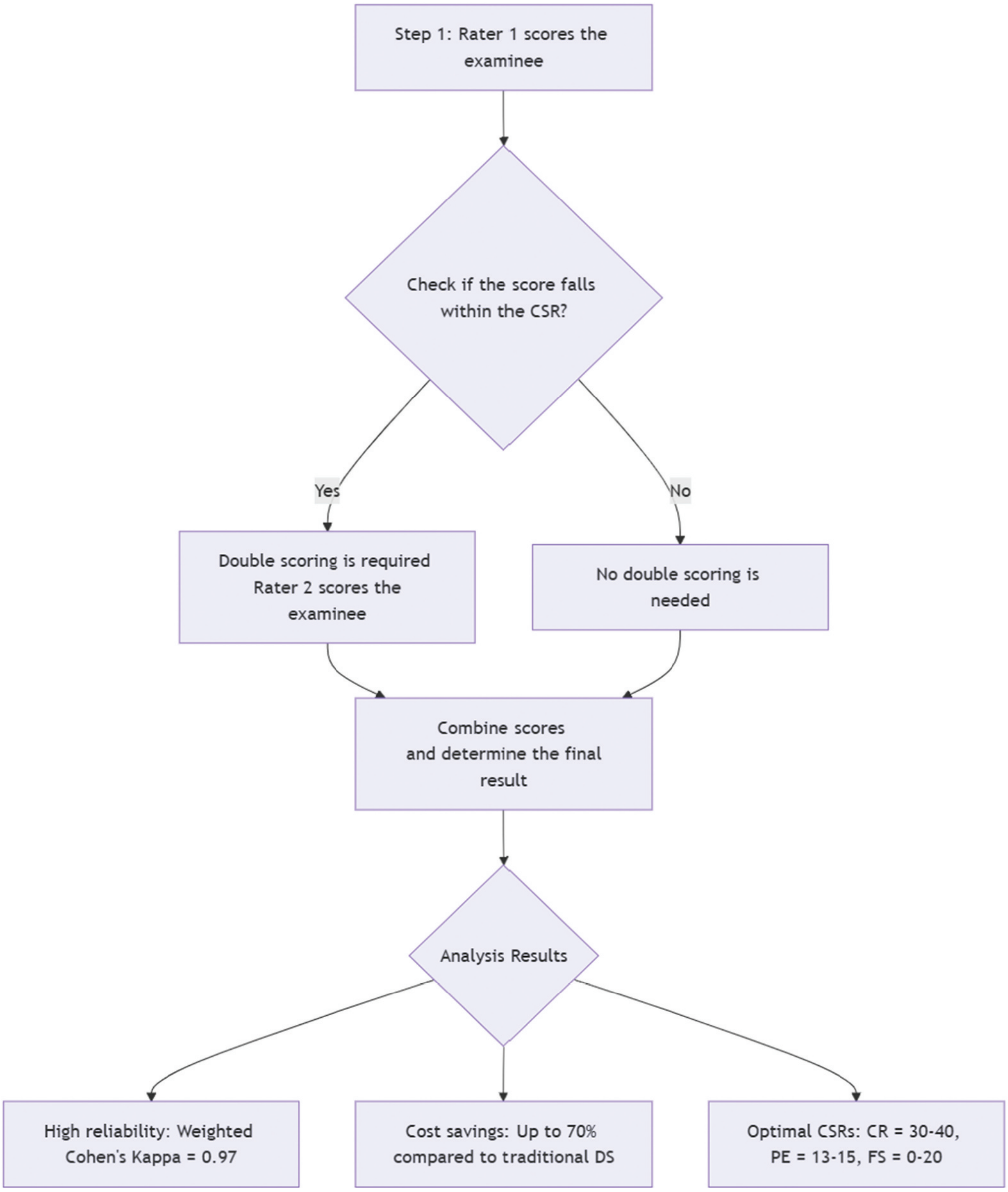

Step 1: Rater 1 scores the examinee
Check if the score falls within the CSR?
Yes
No
Double scoring is required Rater 2 scores the examinee
No double scoring is needed
Combine scores and determine the final result
Analysis Results
High reliability: Weighted Cohen's Kappa = 0.97
Cost savings: Up to 70% compared to traditional DS
Optimal CSRs: CR = 30-40, PE = 13-15, FS = 0-20

Figure 6. Flowchart of targeted double scoring (TDS) decision-making process with analysis results.

refinement through the use of artificial intelligence and machine learning.

While TDS is designed with flexibility and adaptability in mind, its implementation may face challenges such as resistance to change and the need for faculty training. Best practices for adoption include piloting TDS in smaller cohorts before scaling up. The foundational requirement for TDS implementation is the existing use of traditional DS within the OSCE framework. After implementing TDS, the follow-up practice actually reduces the operational burden. Therefore, institutions that already employ traditional DS are well-positioned to adopt TDS, enhancing their assessment processes with minimal adjustments. TDS is aimed at broadening access to efficient, reliable assessment methods, especially in settings where resource constraints might otherwise limit the use of comprehensive evaluation strategies

Central to the discourse presented in this paper is its methodological inspiration, delineating a clear distinction between the generalizability of applied studies' findings and the applicability of the TDS method itself. The core of this paper lies not within the specific outcomes of applying TDS within our study context but in the framework and approach we propose for enhancing the efficiency and reliability of OSCE assessments. This distinction is crucial, as it underscores the potential for TDS to be adapted and applied across a wide range of assessment settings, irrespective of the characteristics or constraints of those environments. The adaptability of the TDS approach signifies a pivotal shift towards more flexible and context-sensitive assessment strategies in medical education. It invites educators and administrators to reconsider the rigidity of traditional scoring systems in favor of methodologies that can be tailored to the unique demands and opportunities of their specific contexts. In doing so, the generalizability of this method extends beyond the confines of our study, offering a versatile tool for improving assessment practices. The emphasis thus shifts towards understanding and leveraging the methodological principles of TDS, encouraging further exploration into its potential applications and modifications to suit diverse educational landscapes.

While TDS offers significant cost savings, its effectiveness may vary depending on the specific context of the OSCE. Extending studies can explore the scalability of TDS in diverse educational settings and its integration with emerging technologies, meaning the minimum to the maximum of all possible scores are assigned to the lower and higher bounds of a CSR. From a statistical perspective, the baseline implies a constant loss (i.e., 40), since $P(\text{CSR}) = 1$, $P(\text{Pass with 1 rating and fail with 2 ratings \& Outside CSR}) = 0$, and $P(\text{Fail with 1 rating and pass with 2 ratings \& Outside CSR}) = 0$. In the study, the costs of baseline could be outperformed by TDS in the first two stations: all $\{L_P, L_F\}$ combinations except {4,000, 4,000} in CR station evidenced the benefits of TDS, while the same conclusion could be found in PE station when {40, 40} and {800, 200} were used. However, the baseline method should be used in the FS station because none of the $\{L_P, L_F\}$ combinations favored TDS. Deciding L_P and L_F may be the first difficulty that researchers face when modeling the data. Further, the way of splitting CSRs into intervals for the cost-effectiveness evaluation also impose challenges: they can be as fine-grained as one point but making the CSRs meaningless, while a wider interval leads to less saving from original costs of rater labors (e.g., if 0–1 and 2–20 are two candidate CSRs for PE station, that scores falling in 2–20 requiring DS results in nearly no savings).

Like other adoptions into assessment, it necessitates a thorough examination of ethical considerations, particularly concerning privacy and data security. As we navigate the integration of technology in educational assessments, it is imperative to ensure that robust measures are in place to protect the confidentiality and integrity of examinee data. This includes adherence to stringent data encryption standards, secure data storage practices, and transparent data handling policies. Additionally, the shift towards remote assessments introduces the need for clear guidelines to prevent unauthorized access and ensure the authenticity of examinee performances. Addressing these ethical concerns is not only critical for maintaining the trust and credibility of the assessment process but also for upholding the dignity and rights of all participants involved.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the National Natural Science Foundation of China [72474004]; National Natural Science Foundation of China for Young Scholars [72104006, 62407002]; Natural Science Foundation of Fujian Province of China [2021J01981]; National Social Science Foundation of China [21BTJ036]; Key Laboratory of Digital Educational Publishing Technology and Standards and Digital Education Research Institute of the Peoples Education Press [RJB0623001]; Peking University Health Science Center [BMU2021YJ010]; Peking University Health Science Center Medical Education Research Funding Project [2023YB24].

ORCID

Zhihui Fu http://orcid.org/0000-0002-8525-0847
Yuhong Wu http://orcid.org/0009-0000-7799-7767
Lingling Xu http://orcid.org/0000-0001-7112-2134
Fen Cai http://orcid.org/0000-0002-6734-2860
Ren Liu http://orcid.org/0000-0002-6708-4996
Zhehan Jiang http://orcid.org/0000-0002-1376-9439

References

[1] McManus IC, McLeod K. Performance assessment in medical education. Oxford: Oxford University Press; 2017.

[2] Bruno GM, Shapiro JA, Holmboe ES. The use of OSCEs to assess competency in medical education. Acad Med. 2018;93(4):523–527.

[3] Bond CF, Titus LJ. Social facilitation: a meta-analysis of 241 studies. Psychol Bull. 1983;94(2):265–292. doi: 10.1037/0033-2909.94.2.265

[4] Aiello JR, Douthitt EA. Social facilitation from Triplett to electronic performance monitoring. Group Dyn: Theor Res Pract. 2001;5(3):163–180. doi: 10.1037/1089-2699.5.3.163

[5] Brown C, Ross S, Cleland J, et al. Money makes the (medical assessment) world go round: the cost of components of a summative final year Objective Structured Clinical Examination (OSCE). Med Teach. 2015;37(7):653–659. doi: 10.3109/0142159X.2015.1033389

[6] Jiang Z, Shi D, Distefano C. A short note on optimizing cost-generalizability via a machine-learning approach. Educ Psychol Meas. 2021;81(6):1221–1233. doi: 10.1177/0013164421992112

[7] Jiang Z, Ouyang J, Li L, et al. Cost-effectiveness analysis in performance assessments: a case study of the objective structured clinical examination. Med Educ Online. 2022;27(1):2136559. doi: 10.1080/10872981.2022.2136559

[8] Wu H, Chen B, Wang H, et al. Remote scoring for an OSCE: a reliable and valid assessment tool for medical education. 2020.

[9] Arnold MK, Cianciolo AT, Audia PF, et al. Evaluation of remote scoring for a medical school clinical skills assessment. Clin Teach. 2019;16(6):466–471.

[10] Majumder MA, Kumar A, Krishnamurthy K, et al. An evaluative study of objective structured clinical examination (OSCE): students and examiners perspectives. Adv Med Educ Pract. 2019;10:387–397. doi: 10.2147/AMEP.S197275

[11] Chong L, Taylor S, Haywood M, et al. The sights and insights of examiners in objective structured clinical examinations. J Educ Eval Health Prof. 2017;14:34. doi: 10.3352/jeehp.2017.14.34

[12] Kim HS, Lee YH, Kim MJ, et al. Development and evaluation of an objective structured clinical examination for assessing the clinical performance of medical students. BMC Med Educ. 2017;17(1):237.

[13] Brand HS, Schoonheim-Klein M. Is the OSCE more stressful? Examination anxiety and its consequences in different assessment methods in dental education. Eur J Dent Educ. 2009;13(3):147–153. doi: 10.1111/j.1600-0579.2008.00554.x

[14] Alfaris EA, Alenazi YR, Alshagga MA, et al. Students' perceptions towards Objective Structured Clinical Examinations (OSCE) at a new medical school in Saudi Arabia. Int J Health Sci (Qassim). 2018;12(1):3.

[15] Park BK, Kim MR, Cho YJ, et al. A comparative analysis of examinee's perception on Objective Structured Clinical Examination (OSCE) between internal and external examiners. J Educ Eval Health Prof. 2017;14:34. doi: 10.3352/jeehp.2017.14.34

[16] Bell CA, Jones ND, Qi Y, et al. Strategies for assessing classroom teaching: examining administrator thinking as validity evidence. Educ Assess. 2018;23(4):229–249. doi: 10.1080/10627197.2018.1513788

[17] Williamson DM, Xi X, Breyer FJ. A framework for evaluation and use of automated scoring. Educ Meas: Iss Pract. 2012;31(1):2–13. doi: 10.1111/j.1745-3992.2011.00223.x

[18] Wiggins G. The case for authentic assessment. Pract Assess, Res, Eval. 1990;2(1):2. doi: 10.7275/ffb1-mm19

[19] Finkelman M, Darby M, Nering M. A two-stage scoring method to enhance accuracy of performance level classification. Educ Psychol Meas. 2008;69(1):5–17. doi: 10.1177/0013164408322025

[20] Sinharay S, Johnson MS, Wang W, et al. Targeted double scoring of performance tasks using a decision-theoretic approach. Appl Psychol Meas. 2023;47(2):155–163. doi: 10.1177/01466216221129271