Vpliv časovnih in težavnostnih diferencialnih uteži na izboljšanje kazalcev kakovosti OSKI

Effects of differential time and difficulty weighting on the improvement of OSCE quality metrics

Avtor / Author

Matic Mihevc[1,2], Klara Masnik[3], Tadej Petreski[4,5], Nejc Pulko[6], Sebastjan Bevc[4,5]

Ustanova / Institute

[1]Zdravstveni dom Ljubljana, Inštitut za raziskave in razvoj osnovnega zdravstva, Ljubljana, Slovenija; [2]Univerza v Ljubljani, Medicinska fakulteta, Katedra za družinsko medicino, Ljubljana, Slovenija; [3]Univerzitetni klinični center Maribor, Oddelek za očesne bolezni, Maribor, Slovenija; [4]Univerzitetni klinični center Maribor, Klinika za interno medicino, Oddelek za nefrologijo, Maribor, Slovenija; [5]Univerza v Mariboru, Medicinska fakulteta, Katedra za interno medicino, Maribor, Slovenija; [6]Univerzitetni klinični center Maribor, Klinika za interno medicino, Oddelek za hematologijo in hematološko onkologijo, Maribor, Slovenija;

[1]Health Centre Ljubljana, Primary Healthcare Research and Development Institute, Ljubljana, Slovenia; [2]University of Ljubljana, Medical Faculty, Chair of Family Medicine, Ljubljana, Slovenia; [3]University Medical Centre Maribor, Department of Ophthalmology, Maribor, Slovenia; [4]University Medical Centre Maribor, Clinic for Internal Medicine, Department of Nephrology, Maribor, Slovenia; [5]University of Maribor, Faculty of Medicine, Chair of Internal Medicine, Maribor, Slovenia; [6]University Medical Centre Maribor, Clinic for Internal Medicine, Department of Haematology, Maribor, Slovenia;

Ključne besede:
OSKI, klinične veščine, ocenjevanje vrstnikov, merjenje kakovosti, obteževanje, osamelci

Key words:
OSCE, clinical skills, peer assessment, quality metrics, weighting, outliers

Članek prispel / Received
22. 10. 2020
Članek sprejet / Accepted
18. 5. 2022

Izvleček

Namen: Objektivni strukturirani klinični izpit (OSKI) je postal vodilna metoda za ocenjevanje izvedbe kliničnih veščin. Njegova pogosta težava je neskladje med oceno na ocenjevalnem obrazcu in splošno oceno ocenjevalca. V preteklosti neskladij niso uspešno reševali z obteževanjem posameznih korakov. Sistematičnega diferencialnega obteževanja, ki podeljuje točke glede na izračun časovnih in težavnostnih uteži, še niso raziskali. Namen študije je bil: a) tvoriti formule za izračun težavnostnih in časovnih uteži, b) tvo-

Abstract

Purpose: Objective structured clinical examination (OSCE) has become a leading method for assessing clinical skills. A common problem observed in OSCE is the discrepancy between checklist scores and global scale scores. To reduce this discrepancy, checklists have been unsuccessfully weighted. However, differential weighting based on a combination of time and difficulty weighting had not been investigated. The goals of our study were to: (a) develop new formulas for time and difficulty weights, (b) design an experimental

https://doi.org/10.18690/actabiomed.230, CC BY 4.0

Naslov za dopisovanje / Correspondence
asist. Matic Mihevc, dr. med.
Inštitut za raziskave in razvoj osnovnega zdravstva, Metelkova ulica 9, 1000 Ljubljana
E-pošta: mihevc.matic@zd-tr.si

riti eksperimentalni obrazec z upoštevanjem izračunanih uteži in c) oceniti uspešnost metode pri izboljšanju kazalnikov kakovosti OSKI.

Metode: *Najprej smo tvorili formule za izračun časovnih in težavnostih uteži posameznih korakov. Nato smo s pomočjo uteži obtežili posamezni korak v eksperimentalnem obrazcu. Kontrolni obrazec so neodvisno od raziskave obtežili klinični mentorji. Uspešnost obrazcev smo preverili med rednim izpitom OSKI za študente 3. letnika. Študente (n = 55) je z obema ocenjevalnima obrazcema in splošnim obrazcem ocenilo deset ocenjevalcev. Rezultate smo analizirali s t-testom za odvisna vzorca, enostavno linearno regresijo in psihometričnim testiranjem.*

Rezultati: *V primerjavi s kontrolnim obrazcem smo na eksperimentalnem obrazcu ugotovili višjo povezavo med oceno na ocenjevalnem obrazcu in splošno oceno ($r = 0{,}622$, $p < 0{,}001$ vs. $r = 0{,}496$, $p < 0{,}001$) in pomembno nižje rezultate ($p < 0{,}001$).*

Zaključek: *Ugotovili smo, da diferencialno obteževanje korakov s časovnimi in težavnostnimi utežmi lahko izboljša kazalnike kakovosti OSKI in zmanjša neskladje med oceno na ocenjevalnem obrazcu in splošno oceno.*

checklist considering the calculated weights, and (c) evaluate the effectiveness of the method in improving OSCE quality metrics.

Methods: *First, we created formulas for time and difficulty weights for each checklist item. Second, we formulated an experimental checklist considering the calculated weights of each item. Simultaneously, a control checklist was independently weighted by OSCE board members. Third, we compared the experimental and control checklists during the OSCE, wherein students (n=55) were graded by ten assessors using both checklists and the holistic global rating form. Finally, we performed statistical analysis with paired samples t-test, simple linear regression, and scale reliability statistics.*

Results: *A higher correlation coefficient between the checklist and the holistic scoring form ($r=0.622$, $p<0.001$ vs. $r=0.496$, $p<0.001$), and significantly lower checklist scores ($p<0.001$), were obtained with the experimental checklist than the control checklist.*

Conclusion: *Assigning time and difficulty differential weights to checklist items improved OSCE quality metrics and reduced the discrepancy between global and checklist scores.*

INTRODUCTION

Clinical examination is a hallmark of medical practice. Several assessment methods have been proposed to test the competence of medical students in clinical examination. Objective structured clinical examination (OSCE) has become a leading method for assessing clinical skills in undergraduate education, progress evaluation, and licensing examinations (1). The OSCE was developed to assess clinical competencies or skills that cannot be sufficiently assessed through other methods, such as oral, written, or computer-based methods (2). A well-designed OSCE is considered objective, reliable, and valid, but it should be well planned and implemented, because it is resource intensive (1, 3). OSCE assessors typically use two types of assessment methods. First, analytical assessment uses a checklist

scale to assess each expected action. It may be scored on a 3-point scale (performed, not performed correctly, or not performed) or a 5- to 7-point scale that also allows assessors to rate the quality of the action. Second, holistic assessment is used to assess the entire process (e.g., empathy, degree of coherence, verbal expression, and nonverbal expression) and is more appropriate for assessing skills in which quality is highly important (4). A group of experts who determine the difficulty and relevance of the test items should agree on the criterion standards before the test (5–6). Several quality indicators should be considered to ensure an objective and valid OSCE (Table 1) (7–11). However, the choice of indicator depends on the design of the OSCE. If both analytical and holistic assessments are used, their results should show strong agreement. This agreement is usually assessed with multiple metrics: a correlation coefficient greater than 0.7, a coefficient of determination greater than 0.5, and inter-grade discrimination one-tenth the total score are considered to reflect an appropriate relationship between scales (7, 9, 11).

Previous research has suggested that using complex, expert-derived weighting schemes may not improve the quality of assessment measurement (12, 13). Several weighting methods have been tested. Initially, dichotomous scoring algorithms (using scores of 1 for all correct options and 0 for one or more missing options) and partial crediting algorithms (assigning partial credit for as many correct answers as possible) have been proposed. However, neither method has been shown to increase the reliability of the results (13). Consequently, differential weighting was introduced (weighting of each option according to its perceived relative clinical importance). Differential weighting has been found to perform better than dichotomous or partial-credit algorithms because more information regarding the examinees' abilities is included in the assessment. However, in most cases, the time spent on weighting has not translated into better test scores (13).

A general problem in previous research is the lack of systematic consideration of a combination of the difficulty and the time spent on an item. Furthermore, no research has assessed validity in contexts in which OSCE checklists are revealed to students. In our OSCE context, the OSCE checklists are disclosed to students as part of their study materials, whereas the scoring algorithms are not disclosed. Therefore, in most cases, students score high on the OSCE exam. Recently, we have found that the correlation between the checklist and the holistic global scores in our OSCE is low, thus resulting in low OSCE quality metrics. We used checklists that were weighted differentially according to OSCE

Table 1. *Interpretation of OSCE quality metrics*

Metric	Brief description	High quality	Low quality
Cronbach's alpha	Measure of internal consistency	>0.7	<0.7
Coefficient of determination (R^2)	Proportional change in checklist score due to a change in global score	>0.5	<0.5
Inter-grade discrimination	Average increase in checklist score equal to a one grade increase in global score	1/10 of the total score	Above or below 1/10 of the total score
Number of failures	Number of students failing per OSCE station	Individually	Individually
Between-group variation	Proportion of variation in the checklist score arising from student performance rather than changes in environment, location, or assessor bias	<30%	>40%

Formula 1. *Item time factor*

$$\mathrm{ITF} = \frac{\frac{1}{n} \times \sum_{i=1}^{n} SIPTi}{\mathrm{PAL}} \times 100$$

Legend: ITF, item time factor; SIPT, single item performance time; n, number of measured SIPT; PAL, protocol allocated time.

Formula 2. *Item difficulty factor*

$$\mathrm{IDF} = 1 + \frac{\mathrm{IPI} + \mathrm{OI}}{\mathrm{TIP}}$$

Legend: IDF, item difficulty factor; IPI, number of items performed incorrectly; OI, number of items omitted; TIP, total number of items performed.

Formula 3. *Final item points*

$$\mathrm{FIP} = \frac{\mathrm{ITF} \times \mathrm{IDF}}{\sum_{i=m}^{n} a_i = (\mathrm{ITFxIDF})_m + (\mathrm{ITFxIDF})_{m+1} + \ldots + (\mathrm{ITFxIDF})_{n-1} + (\mathrm{ITFxIDF})_n} \times 100$$

Legend: FIP, final item points; ITF, item time factor; IDF, item difficulty factor.

board members' subjective opinions regarding the importance of each item. No systematic method was used to weight the OSCE items. To improve OSCE quality metrics, we developed a new approach for weighting OSCE checklists.

The goals of our study were to (a) develop new formulas for time and difficulty weights, (b) design an experimental checklist considering the calculated weights, and (c) evaluate the effectiveness of the method in improving OSCE quality metrics.

MATERIAL AND METHODS

Experimental checklist formation

We created a new experimental OSCE checklist for the cardiovascular system (CVS) examination. The checklist consisted of 36 items (Supplementary Table). The allotted time to complete the protocol was 14 minutes. The allotted time was set by OSCE board members not involved in this study. We differentially weighted the experimental checklist on the basis of the calculated time and difficulty weights. A total of 100 points (percentages) were targeted in the experimental checklist for ease of scoring.

First, we defined time weight with the item time factor (ITF). A total of 18 peer tutors performed each checklist item (Supplementary Table) perfectly, with no time limit per individual item. We measured the time required for each item and labelled the result as a single item performance time (SIPT). We calculated the ITF by dividing the average SIPT of the 18 peer tutors by the protocol allocated time (PAL) and multiplying the coefficient by 100 to obtain percentages (Formula 1).

Second, a formula for item difficulty weight was developed on the basis of a study by Shulruf et al., who have found that item difficulty successfully discriminates between passing and failing students with borderline OSCE scores (14). We defined the difficulty weight with the item difficulty factor (IDF). We analysed previous OSCE evaluation reports (2012–2016) identifying the number of correct, incorrect, and omitted performance results for each item. We set the default IDF to 1 and increased it according to the coefficient between the sum of incorrect and omitted items relative to the total number of items performed (Formula 2).

Finally, we defined the final item points (FIP) by multiplying the ITF by the IDF of an individual item relative to the sum of the products of ITF and IDF for all individual items. We then multiplied the result by 100 to obtain percentages (Formula 3). We rounded FIP to the nearest half-point for easier scoring.

Control checklist formation

A control OSCE checklist was independently formulated by the OSCE board members, who weighted each item according to their subjective opinion regarding the importance of each item.

Holistic scoring form

We used a holistic scoring form based on an article by Hodges et al. (4). The form was translated into Slovenian through a backward and forward process. Finally, the form was tested by users (peer tutors) for clarity and feasibility. A form consisted of four categories and an overall global score. Examiners assessed students' communication skills, degree of coherence in the examination, technique of the examination, and ability to report findings correctly. Assessors were also asked to provide an overall assessment of the knowledge and skills demonstrated in the examination. We used a 5-point Likert scale for each category and overall assessment (1 = clear fail, 5 = excellent). Cronbach's alpha for this new instrument was 0.79.

OSCE context and assessor training

In the introduction, we described the unique context of our OSCE. In recent years, we have recognised peer teaching of clinical skills as an important part of lifelong learning among physicians (15). Consequently, we have replaced faculty-led clinical skills courses with student-led courses with mentorship by faculty members. Clinical skills peer teachers are required to complete rigorous clinical skills training conducted by faculty members. The quality of teaching is also ensured by providing OSCE checklists to students and peer teachers as part of their study materials. The courses are therefore strictly based on the OSCE checklists. They last for 1 month and end with a final OSCE exam in which students' knowledge is assessed by peer teachers at five stations, including the CVS examination. However, the checklist scores are not shared with peer assessors until the OSCE exam and are known only to the OSCE board members. In addition, quality control of the peer assessment is performed by the faculty members at the OSCE site. For the purposes of this study, we provided additional training to the peer assessors 2 days before the OSCE. Each item in the CVS checklist was demonstrated and discussed with the faculty members. In addition, the categories of the global rating form were discussed and standardised.

Comparative study design

The experimental checklist was tested in a comparative study during the regular OSCE for 3rd-year medical students in November 2017. Students' performance on the CVS examination was assessed with the control checklist, the experimental checklist, and the holistic scoring form. First, the assessors evaluated the examination performance by using the control and experimental checklists simultaneously. Before summing the scores for both checklists, they completed the holistic scoring form. Second, they summed scores for both checklists to eliminate study bias. The students signed a written informed consent form stating their agreement with the study and the use of their results for research purposes. The institutional ethics committee approved the study. The students' personal data were protected in compliance with the Law on Personal Data Protection.

Statistical analysis

Statistical analysis was performed in the SPPS statistical package, version 22 (SPSS Inc., Chicago, IL, USA) for Windows, IBM. Results are presented as means±standard deviations or percentages. Paired samples t-test was used to compare checklist results, and a simple linear regression model and scale reliability analysis (Cronbach's alpha) were used to assess checklist quality. Cook's distance was calculated to identify students with outlying results. A threshold of 0.07 (four/number of observations) was used to detect influential outliers (16). Nevertheless, we conducted the final analysis with outliers included to avoid research bias. The threshold for statistical significance was set at $p<0.05$.

RESULTS

Experimental checklist

Peer tutors performed the CVS examination within an average of 814.4±4.5 seconds. Most of the time was spent on the item "palpation of the peripheral arterial pulses of the leg" (97.4±23.1 s). In contrast, the least amount of time was spent on the item "gaining patient consent" (2.1±0.5 s).

The most difficult item to perform correctly in previous OSCEs was "deep and superficial palpation of the abdomen", with an IDF of 1.43. The easiest items, with an IDF of 1.00, were those associated with communication with the patient: "introduction to the patient", "gaining the patient's consent", and "asking the patient to undress".

The highest FIP was calculated for the item "palpation of peripheral arterial pulses of the leg" (11.0 points). The lowest FIPs (0.5 points) were given for the items "disinfection of the hands", "introduction to the patient", "explanation of the nature of the examination", and "gaining patient's consent". Data for selected individual items in the experimental and control checklists are shown in Table 2. More detailed data are presented in the Supplementary Table.

Comparative study

Students completed the CVS examination within 818.0±36.0 s on average. The mean checklist score was significantly lower for the experimental checklist than the control checklist (89.9±6.8% vs. 92.±5.9%, p<0.001). The mean overall global score was 4.2±0.5. The highest scoring component of the holistic assessment form was the ability to report findings (4.4±0.6), followed by communication skills (4.3±0.6), coherence in examination (4.2±0.7), and examination technique (4.2±0.6).

Table 2. *Selected results for individual items*

Protocol item	SIPT Mean±SD (s)	ITF	IDF	Experimental checklist FIP (%)	Control checklist FIP (%)
Disinfection of the hands	6.6±1.9	0.80	1.01	0.5	3.0
Palpation of the cervical lymph nodes	32.9±4.7	4.82	1.23	4.5	3.0
Assessment of the central venous pressure	49.2±9.2	6.05	1.14	6.0	3.0
Palpation of the precordium	50.5±17.5	6.20	1.17	6.0	8.0
Auscultation of the heart at five sites with the membrane	51.1±11.9	6.28	1.34	7.0	8.0
Deep and superficial palpation of the abdomen	43.5±5.7	5.34	1.43	6.5	6.0
Palpation of peripheral arterial pulses of the leg	97.4±23.1	11.96	1.09	11.0	8.0
Auscultation of the back side of the thorax	34.0±4.7	4.18	1.15	4.0	2.0

Legend: SIPT, single item performance time; SD, standard deviation; ITF, item time factor; IDF, item difficulty factor; FIP, final item points.

The correlation coefficient between the experimental checklist score and the global score indicated a moderate correlation, r=0.622, p<0.001, which explained 38.7% of the variation in the experimental checklist score (R2) (Figure 1a). The inter-grade discrimination was 7.96.

The correlation coefficient between the control checklist score and the global score showed a low correlation, r=0.496, p<0.001, which explained 24.6% of the variation in the control checklist score (R2) (Figure 1b). The inter-grade discrimination was 5.50.

The correlation coefficient between the control and experimental checklist scores indicated a high correlation, r=0.868 p<0,001 (Figure 1c).

Table 3. *Outliers in the experimental and control checklists*

Checklist	Checklist score (%)	Global score	Cook's distance
Experimental	69.0	4	0.53
Control	80.0	3	0.12
	84.0	3	0.10
	78.0	3	0.12
	97.0	4	0.11
	82.0	5	0.30

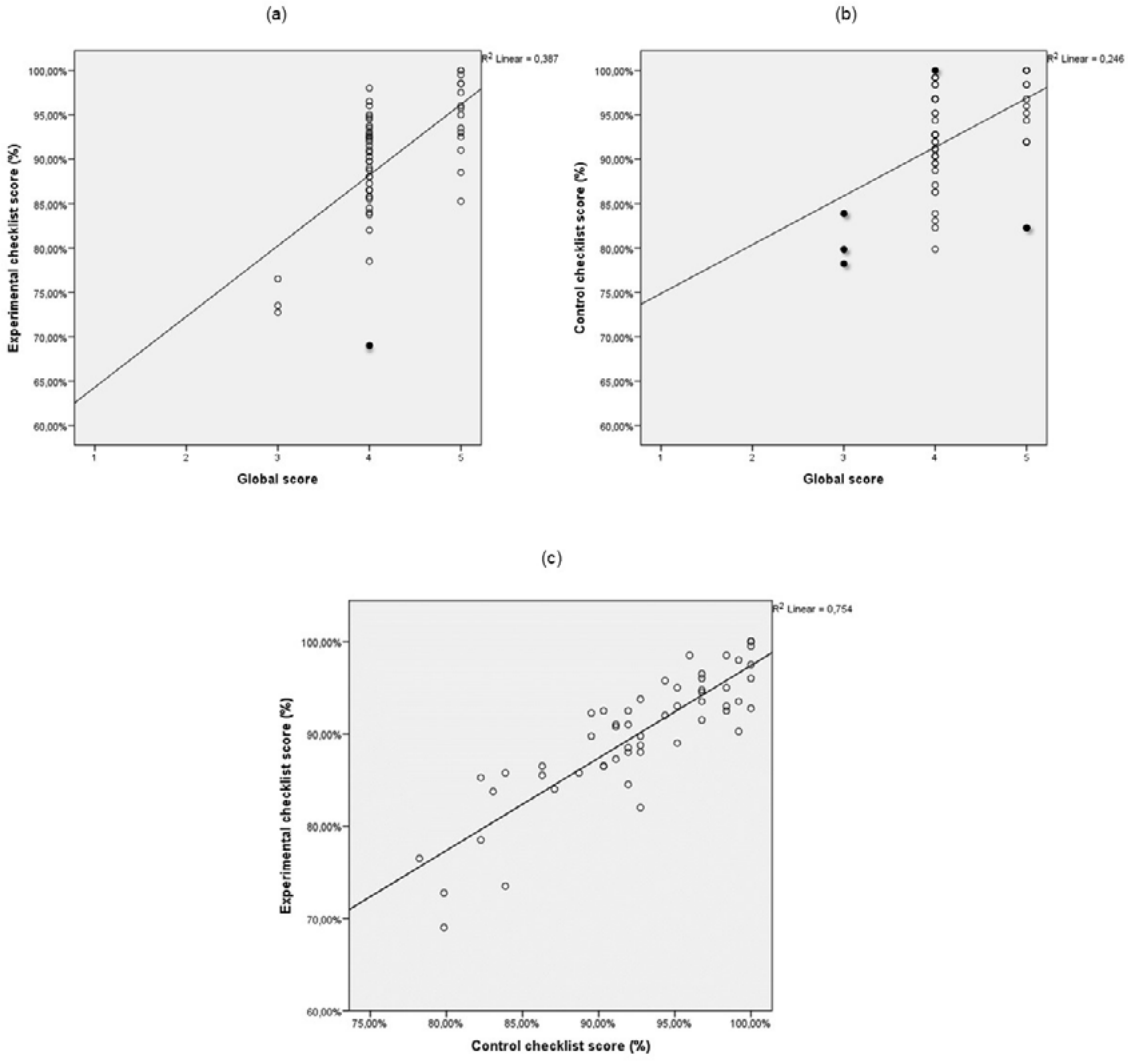

Figure 1. *Linear regression between the (a) experimental checklist score and global score, (b) control checklist score and global score, and (c) experimental and control checklist score. Outliers are coloured black.*

One outlier was identified in the experimental checklist, and five outliers were identified in the control checklist (Table 3).

DISCUSSION

The evidence from this study suggests that differential weighting of a 3-point checklist according to time and difficulty may improve OSCE quality metrics and reduce the discrepancy between the analytic checklist score (what was done) and the holistic global score (how well it was done).

The experimental checklist, compared with the control checklist, achieved a higher correlation coefficient between the checklist and global scores, a higher determination factor, a more desirable inter-grade discrimination value, and significantly lower checklist scores. The criteria for a high quality OSCE station (Table 1) were nearly met for the experimental checklist, whereas the control checklist did not meet them.

The intercorrelation coefficient of the checklists was another indicator strongly suggesting that the observed differences between the experimental and control checklists were associated with the weighting schemes rather than other factors (Figure 1c). Previous studies advising against the use of weighting schemes have reported that scores derived from different weighting methods are often correlated, with an intercorrelation coefficient of 0.9 or higher, whereas the intercorrelation coefficient in our study was less than 0.9 (13).

In general, weighting OSCE checklist items or components is perceived as not worth the effort, and experts' time is believed to be better spent on developing new tests (12, 13, 17, 18). However, several examination boards and academic institutions still apply differential scoring weights. In addition, the 2014 International Test Commission guidelines suggest using weights when a rationale to do so is present (19). Therefore, the effectiveness of item weighting is perceived differently across OSCE contexts and by different researchers.

In contrast, another perspective on weighting drove our research. Several previous studies have described the successful weighting of exam subcomponents (e.g., medical history taking, physical examination, communication skills, and interpersonal skills) to produce sound and valid overall scores. They have found that faculty weighting is insufficient to produce valid results. Instead, they have proposed combining faculty opinions with weights considering the reliability, variance, and the association (covariance) of subcomponents with one another (20–23).

Although the results of our study are promising, they should be interpreted with caution, and limitations should be addressed. First, the correlation coefficient between the experimental checklist and the holistic global score in our study was in the range of 0.6–0.7, whereas the control checklist showed even worse performance. However, a perfect correlation was not expected, because analytic and holistic scales do not precisely capture the same aspects of the examination. Therefore, a correlation coefficient of 0.7 (explaining 50% of the observed variance) was considered satisfactory (7–11).

Previous research has suggested that several factors may explain the discrepancy between scores. One group of factors is outliers, which can significantly affect the correlation between global and checklist scores, as was the case in our study. Three main groups of OSCE outliers are described: (a) students who perform poorly and achieve a checklist score close to zero, (b) students who perform well and achieve a low global score, and (c) students who are graded unreliably by the assessors (7, 24). In our study, we found one outlier in the experimental checklist and five outliers in the control checklist. Figure 1a (experimental checklist) shows a significant outlier in the case of a student who received a high global score but did not score highly on the checklist, owing to poor time management: the student did not complete the entire examination protocol in the allocated time and did not achieve the remaining points. In contrast, Figure 1b (control checklist) reveals four outliers in cases of students who had high control checklist scores but low overall global scores. These students rushed through or skipped items that required more time to complete but were awarded fewer points on the overall holistic form. In the experimental checklist, those items were worth more points because of time weighting, and students were penalised more for not performing them. Therefore, no outliers of this type were found. Rushing or skipping behaviour could be further discouraged by allocating students more time for performing the OSCE station. To avoid research

bias, the data are presented with outliers included.

Second, the correlation may also be affected by the unreliable grading of assessors who assign global scores according to checklist results (10). We attempted to eliminate this effect by prohibiting assessors from summing checklist scores before completing the holistic global score checklist.

Third, several factors among groups of assessors, such as the stringency-leniency effect, the halo effect, the gender effect, and others, may affect the correlation (24). In another conference paper, we addressed this issue and found that our assessors were susceptible to the stringency-leniency effect, whereas no other effects were found (25). Thus, the stringency-leniency effect might have negatively or positively affected the correlation coefficient in our study.

Fourth, we used clinical skills peer teachers as assessors, who might have lacked sufficient experience to provide a satisfactory global score (26, 27). However, owing to faculty oversight, we do not consider this aspect to be a limitation (28). Furthermore, a recent study by Donohoe et al. has indicated that neither clinical experience nor relevant content experience independently predicted the overall grade assigned in the OSCE (29).

Finally, we used a thorough (long) checklist for our OSCE. Previous research has suggested that restricting the checklist to clinically relevant features leads to better accuracy and better psychometric indices (30). We do not disagree with previous findings. However, the checklist was revealed to students before this research was conducted and served as a learning tool for them. Few studies have examined the effect of OSCE checklist disclosure on the scores obtained on the OSCE. Chae et al. have found no significant difference in overall scores before or after disclosure between the experimental and control groups (31). We noted that a "side effect" of prior OSCE checklist disclosure was high OSCE checklist scores (>90%). Therefore, we were unable to fully assess the effect of weighting items with item difficulty, whereas other studies have shown that students with borderline pass scores are most likely to benefit (14).

CONCLUSION

In conclusion, weighting may sometimes perform favourably, particularly in an OSCE context such as ours. Implementation of time and difficulty weighting is a novel and systematic approach that, to our knowledge, has not been previously proposed. In addition to improving OSCE quality metrics, time weighting successfully penalised students who rushed through the OSCE station or skipped items that required more time for successful completion of the OSCE station. However, because our method is time-consuming, we advise others working in similar OSCE contexts to first use simpler methods to improve the metrics, such as using simplified OSCE checklists, splitting long OSCE stations into multiple short OSCE stations, and reducing the number of students per assessor to avoid leniency effects.

ACKNOWLEDGEMENTS

We thank the clinical skills peer tutors and the students who participated in the study, and extend special thanks to the Faculty of Medicine, University of Maribor, for supporting peer teaching of clinical skills. Finally, we thank Tamara Serdinšek for the analysis of previous OSCE results.

Supplementary Table. *Differential weights in the control and experimental checklists*

Step	Item	Control checklist weight (%)	Experimental checklist weight (%)
1	Disinfect your hands.	3.0	0.5
2	Introduce yourself to the patient.	2.0	0.5
3	Explain the nature of the examination.	2.0	0.5
4	Gain patient consent.	2.0	0.5
5	Ask the patient to undress.	2.0	0.5
6	Determine the general impression of the patient (consciousness, orientation, cyanosis).	2.0	1.5
7	Adjust the backside of the table to 45°. Ask the patient to lie down.	2.0	1.0
8	Inspect the palms and the fingers (clubbed fingers, tar stains, bilateral capillary pulsations on the nails).	3.0	2.0
9	Palpate and compare the radial pulses (pulse volume, symmetry), using at least two fingers for palpation.	3.0	1.0
10	Evaluate the heart rate (beats/min), palpating the pulse for at least 30 s.	3.0	4.0
11	Inspect the face (facies mitralis, xanthelasmas, arcus senilis) and conjunctiva.	3.0	2.0
12	Inspect the oral cavity and pharynx (teeth, tonsils, hydration, central cyanosis), using a spatula and light.	2.0	3.0
13	Inspect and palpate the neck; palpate cervical lymph nodes (groups of lymph nodes: submental, sublingual, submandibular, anterior and posterior, suboccipital, supraclavicular); and determine the carotid pulse.	5.0	5.5
14	Perform central venous pressure assessment (cm H_2O).	3.0	6.0
15	Inspect the thorax (scars, precordial pulsations, asymmetry, hyperinflation).	2.0	1.5
16	Palpate the precordium (apex location, amplitude, size in cm^2).	8.0	6.0
17	Auscultate the heart (rhythm, heart sounds, heart murmurs), and palpate the radial pulse while auscultating.	5.0	4.5
18	Auscultate the apex.	2.0	1.0
19	Auscultate the 5^{th} right ICS parasternal.	2.0	1.0
20	Auscultate the 2^{nd} left ICS parasternal.	2.0	1.0
21	Auscultate the 2^{nd} right ICS parasternal.	2.0	1.0
22	Auscultate the 3^{rd} left ICS parasternal (Erb's point).	2.0	1.0
23	Auscultate the sites mentioned above with a bell.	3.0	3.5
24	Auscultate the carotid arteries (patient should be asked to hold the breath).	2.0	1.5
25	Before the examination of the abdomen, adjust the back section of the examination table to 15–20°.	2.0	2.0
26	Inspect (symmetry, level) and palpate the abdomen (superficial and deep palpation, palpate the liver, spleen, and other masses, aortic pulsations, painful sites).	8.0	13.5
27	Auscultate the abdomen (bruits of the aorta, renal and iliac arteries).	3.0	5.0
28	Inspect the legs (oedema, skin colour, nails, varices), and assess capillary refill (hold the tips of both big toes for 5 s).	3.0	3.5
29	Detect pretibial and ankle oedema (press the pretibial region/ankles and hold for 15 s).	2.0	4.5
30	Palpate posterior calves for deep vein thrombosis detection.	3.0	1.0
31	Palpate the femoral arteries, popliteal arteries, posterior tibial arteries, and dorsalis pedis arteries. Auscultate the femoral arteries.	8.0	11.0
32	Perform percussion of the back side of the thorax (symmetrically left and right).	2.0	3.5
33	Auscultate the back side of the thorax (symmetrically left and right). The patient should breathe through the open mouth.	2.0	4.0
34	Explain the findings to the patient.	2.0	1.0
35	Thank the patient for cooperating.	2.0	0.5
36	Disinfect your hands.	2.0	0.5

Legend: ICS, intercostal space.

REFERENCES

1. Hastie MJ, Spellman JL, Pagano PP, Hastie J, Egan BJ. Designing and implementing the objective structured clinical examination in anesthesiology. Anesthesiology. 2014;120(1):196-203.
2. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. Med Teach. 2013;35(9):e1437-46.
3. Barman A. Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singapore. 2005;34(8):478-82.
4. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 2003;37(11):1012-6.
5. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: organisation & administration. Med Teach. 2013;35(9):e1447-63.
6. Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. Med Teach. 2018;40(12):1208-13.
7. 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.
8. Al-Osail AM, Al-Sheikh MH, Al-Osail EM, Al-Ghamdi MA, Al-Hawas AM, Al-Bahussain AS et al. Is Cronbach's alpha sufficient for assessing the reliability of the OSCE for an internal medicine course? BMC Res Notes. 2015;8(1):582.
9. Hejri SM, Jalili M, Muijtjens AM, Van Der Vleuten CP. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci. 2013;18(10):887-91.
10. Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach. 2015;37(12):1106-13.
11. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):69-71.
12. Sandilands DD, Gotzmann A, Roy M, Zumbo BD, De Champlain A. Weighting checklist items and station components on a large-scale OSCE: is it worth the effort? Med Teach. 2014;36(7):585-90.
13. Bordage G, Page G. The key-features approach to assess clinical decisions: validity evidence to date. Adv Health Sci Educ Theory Pract. 2018;23(5):1005-36.
14. Shulruf B, Booth R, Baker H, Bagg W, Barrow M. Using the Objective Borderline Method (OBM) to support Board of Examiners' decisions in a medical programme. J Furth High Educ. 2017;41(3):425-34.
15. Zdravkovic M, Serdinsek T, Sobocan M, Bevc S, Hojs R, Krajnc I. Students as partners: Our experience of setting up and working in a student engagement friendly framework. Med Teach. 2018;40(6):589-94.
16. Bollen KA, Jackman RW. Regression diagnostics:An expository treatment of outliers and influential cases. Sociol Methods Res. 1985;13(4):510-42.
17. Bobko P, Roth PL, Buster MA. The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. Organ Res Methods. 2007;10(4):689-709.
18. Bolger F, Rowe G. The aggregation of expert judgment: Do good things come to those who weight? Risk Anal. 2015;35(1):5-11.
19. International Test C. ITC Guidelines on quality control in scoring, test analysis, and reporting of test scores. Int J Test. 2014;14(3):195-217.
20. Park YS, Lineberry M, Hyderi A, Bordage G, Xing K, Yudkowsky R. Differential weighting for subcomponent measures of integrated clinical encounter scores based on the USMLE Step 2 CS Examination: Effects on composite score reliability and pass-fail decisions. Acad Med. 2016;91(11):S24-S30.
21. Clauser BE, Balog K, Harik P, Mee J, Kahraman N. A multivariate generalizability analysis of history-taking and physical examination scores from the USMLE step 2 clinical skills examination. Acad Med. 2009;84(10):S86-9.
22. Feldt LS. Estimating the reliability of a test battery composite or a test score based on weighted item scoring. Meas Eval Couns Dev. 2004;37(3):184-90.
23. Kane M, Case SM. The reliability and validity of weighted composite scores. Appl Meas Educ. 2004;17(3):221-40.

24. Schleicher I, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B et al. Examiner effect on the objective structured clinical exam - A study at five medical schools. BMC Med Educ. 2017;17(1):71.
25. Mihevc M, Masnik K, Petreski T, Pulko N, Bevc S. Factors influencing accessor's checklist and global scores at OSCE In: Plass H, editor. 22nd Graz Conference: Whereto is medical education going; 2018 Apr 5-7; Maribor, Slovenia: ÖGHD; 2018. p. 29-30.
26. Iblher P, Zupanic M, Karsten J, Brauer K. May student examiners be reasonable substitute examiners for faculty in an undergraduate OSCE on medical emergencies? Med Teach. 2015;37(4):374-8.
27. Khan R, Payne MWC, Chahine S. Peer assessment in the objective structured clinical examination: A scoping review. Med Teach. 2017;39(7):745-56.
28. Rižnik P, Bevc S. The evolution of clinical skills peer teaching at the Faculty of Medicine, University of Maribor. Acta Medico-Biotechnica. 2016;9(2):17-24.
29. Donohoe CL, Reilly F, Donnelly S, Cahill RA. Is there variability in scoring of student surgical OSCE performance based on examiner experience and expertise? J Surg Educ. 2020;77(5):1202-1210.
30. Yudkowsky R, Park YS, Riddle J, Palladino C, Bordage G. Clinically discriminating checklists versus thoroughness checklists: improving the validity of performance test scores. Acad Med. 2014;89(7):1057-62.
31. Chae SJ, Kim M, Chang KH. Can disclosure of scoring rubric for basic clinical skills improve objective structured clinical examination? Korean J Med Educ. 2016;28(2):179-83.