

BMJ Open Objectivity in subjectivity: do students' self and peer assessments correlate with examiners' subjective and objective assessment in clinical skills? A prospective study

A'man Talal Inayah,¹ Lucman A Anwer,^{1,2} Mohammad Abrar Shareef,^{1,3} Akram Nurhussen,¹ Haifa Mazen Alkabbani,¹ Alhanouf A Alzahrani,¹ Adam Subait Obad,¹ Muhammad Zafar,¹ Nasir Ali Afsar¹

To cite: Inayah A'mT, Anwer LA, Shareef MA, *et al.* Objectivity in subjectivity: do students' self and peer assessments correlate with examiners' subjective and objective assessment in clinical skills? A prospective study. *BMJ Open* 2017;7:e012289. doi:10.1136/bmjopen-2016-012289

► Prepublication history and additional material for this paper are available online. To view these files please visit the journal online (<http://dx.doi.org/10.1136/10.1136/bmjopen-2016-012289>).

Received 14 April 2016

Revised 15 March 2017

Accepted 16 March 2017



CrossMark

¹College of Medicine, Alfaisal University, Riyadh, Saudi Arabia

²Mayo Clinic, Rochester, USA

³Mercy St. Vincent Medical Center, Toledo, USA

Correspondence to

Dr Nasir Ali Afsar;
nafsar@alfaisal.edu

ABSTRACT

Objectives The qualitative subjective assessment has been exercised either by self-reflection (self-assessment (SA)) or by an observer (peer assessment (PA)) and is considered to play an important role in students' development. The objectivity of PA and SA by students as well as those by faculty examiners has remained debated. This matters most when it comes to a high-stakes examination. We explored the degree of objectivity in PA, SA, as well as the global rating by examiners being Examiners' Subjective Assessment (ESA) compared with Objective Structured Clinical Examinations (OSCE).

Design Prospective cohort study.

Setting Undergraduate medical students at Alfaisal University, Riyadh.

Participants All second-year medical students (n=164) of genders, taking a course to learn clinical history taking and general physical examination.

Main outcome measures A Likert scale questionnaire was distributed among the participants during selected clinical skills sessions. Each student was evaluated randomly by peers (PA) as well as by himself/herself (SA). Two OSCEs were conducted where students were assessed by an examiner objectively as well as subjectively (ESA) for a global rating of confidence and well-preparedness. OSCE-1 had fewer topics and stations, whereas OSCE-2 was terminal and full scale.

Results OSCE-1 (B=0.10) and ESA (B=8.16) predicted OSCE-2 scores. 'No nervousness' in PA (r=0.185, p=0.018) and 'confidence' in SA (r=0.207, p=0.008) correlated with 'confidence' in ESA. In 'well-preparedness', SA correlated with ESA (r=0.234, p=0.003).

Conclusions OSCE-1 and ESA predicted students' performance in the OSCE-2, a high-stakes evaluation, indicating practical 'objectivity' in ESA, whereas SA and PA had minimal predictive role. Certain components of SA and PA correlated with ESA, suggesting partial objectivity given the limited objectiveness of ESA. Such difference in 'qualitative' objectivity probably reflects experience. Thus, subjective assessment can be used with some degree of objectivity for continuous assessment.

Strengths and limitations of this study

- It is a prospective study of undergraduate medical students.
- All applicable subjective and objective assessment methods were included in a single cohort.
- The data denote a semester-long (approximately 6 months) observation.
- It is a study from a single institution reporting the observations about junior medical students only.
- The long-term follow-up and observation (beyond 6 months) is lacking.

INTRODUCTION

Medical education is evolving constantly. Physicians, deemed as 'competent' health providers, are expected to be self-directed and active lifelong learners nowadays.¹ Thus, there is a shift from duration-based education to competency-based training. Accordingly, medical curricula were revised at many places. This resulted into development of revision of assessment methods to fit the changing trends,² thus ultimately requiring faculty training and development. Taking a clinical history and conducting physical examination remain fundamental skills learnt by medical students. Clinical history taking not only involves asking questions about patients' illness, but it also requires grasping various techniques to effectively and appropriately communicate with the patient and build a good rapport. Similarly, the science of physical examination is an art that involves specific approach and steps which could make a huge difference to patient management. Traditionally, an Objective Structured Clinical Examination (OSCE) is the method of choice to evaluate the clinical skills of

medical students objectively, where they are judged and graded through checklists, for a given set of standardised observable tasks. Despite its objectivity, passing an OSCE does not guarantee how the students would practice in real life. Another limitation of an OSCE is its being labour and resource-intensive, thus limiting its utility to be a frequently conducted activity for learning, evaluation and feedback. Thus, developing alternate assessment methods to monitor the development of self-directed, lifelong learners is pivotal, beginning with the realisation of personal learning needs,³ which in turn leads to the development of a focused list of personalised learning objectives.⁴

Standardised tests may not provide complete insight into the skills of the trainee physician.⁵ Hence, combining them with other assessment techniques such as self-assessment (SA) and peer assessment (PA) may provide a more holistic view potentially leading to a better outcome.⁶ SA is 'the act of judging one's own self and making decisions about the required steps'.⁷ The role of SA has been studied in the field of education.^{8–10} It has been shown to be helpful in improving knowledge acquisition as well as in enhancing morale, motivation, communication and overall performance.¹¹ Similarly, PA has also been established as an effective educational tool.⁷ According to Falchikov, it requires students 'to provide either feedback or grades (or both) to their peers on a product or a performance, based on the criteria of excellence for that product or event which students may have been involved in determining'.¹² PA can also help to improve student participation and promote them to become lifelong learners.¹³

Another qualitative tool could be Examiners' Subjective Assessment (ESA), which relies on global rating of a student for domains such as proficiency and confidence during standardised clinical examinations.² Although such global ratings have shown contrasting accuracy results,^{6 14 15} their utility in assessing medical students still remains understudied.

We hypothesised that due to the focused nature of the course and its assessment, the subjective evaluations should correlate with OSCE scores, thus making it a surrogate marker of outcome while the course is still in progress. To understand objectivity in these subjective tools, we designed this study to explore any relationship between SA, PA, ESA and OSCE scores in a holistic fashion.

METHODS

This prospective cohort study was conducted at Alfaisal University College of Medicine (AU CoM) in Riyadh during fall semester 2013. AU CoM has adopted SPICES model of curriculum, divided into 10 semesters spanning over 5 years. It is designed in spiral fashion, emphasising a gradual 'basic to clinical' shift in themes and training. During semesters 1–3, organ-system blocks are taught with an emphasis on normal structure and function.

The students are also offered parallel running courses of clinical communication skills, history taking and general physical examination. On the other hand, during semesters 4–6, the organ-system blocks are repeated in the similar sequence, emphasising on pathology, microbiology, pharmacology and clinical aspects, with parallel running clinical skill courses integrated with respective organ-system blocks and themes. Semesters 7–10 comprise only of clinical clerkship at affiliated hospitals. All clinical skills courses—from year 1 to 5—are evaluated with OSCEs.

This study focuses on clinical skills course spanning over 18 weeks of that semester and designed for year 2 medical students (n=164) to introduce essentials of clinical history taking and general physical examination. All students, divided into small groups, had a weekly session, spanning over 2 hours. The course was designed with emphasis on hands-on practice of identified sets of skills pertaining to basics of communication, history taking and only general physical examination (including vital signs) in that semester. After a certain number of weeks, a demonstration session would be planned where all students would demonstrate a subset of their skills learnt over preceding weeks in a semi-isolated small group setting. Each student was evaluated by himself/herself as well as 3–5 of his/her peers (completing SA and PA). Further, the course had two OSCEs, one being mid-semester, small scale (three stations, comprising of full history taking, vital signs, general physical examination), and the other at the end of the course (five stations, comprising of two history-taking stations, one vital sign, two general physical examination stations), full scale. The stations were carefully designed to enable unambiguous testing of only the intended skills. Both OSCEs had a single experienced examiner at each station. Apart from the objective assessment, each OSCE also had a concurrent subjective assessment component where the examiner would assign a global performance rating score or ESA to each student (see online supplementary appendix 1). Before each OSCE, the examiners as well as the educators would meet and standardise the grading on the basis of a customised checklist focusing a given task (see online supplementary appendix 1). Thus, two approaches were used to evaluate each student. First, OSCEs were used for objective assessment. Second, there were three subjective assessments which included ESA done by examiners, SA done by the student himself/herself and PA done by the student's peers (figure 1, see online supplementary appendix 2).



Figure 1 The timeline of various student assessments during the course. ESA, Examiners' Subjective Assessment; OSCE, Objective Structured Clinical Examinations; PA, peer assessment; SA, self-assessment.

A short five-point Likert scale questionnaire was used to record SA and PA, ranging from '1, strongly disagree' to '5, strongly agree', which was developed with a focus on patient-centred competencies adopted from Papinczak *et al* (see online supplementary appendix 2).¹⁴ It was distributed to the students during each demonstration session in the course and assessed the following domains: (1) confidence, (2) respectful manner, (3) attentive listening, (4) absence of nervousness, (5) the use of non-technical language, (6) being concise and (7) appearing well-prepared. All students evaluated themselves using the same questionnaire, representing SA. Simultaneously, each student was also evaluated by a random selection of 3–5 peers on the same parameters as described above, constituting PA. Paper questionnaires were collected immediately at the end of the session by the supervising instructor. All subjective evaluations were part of the multi-faceted approach of the course evaluation and hence did not require additional consent from individual students. The student IDs were used to identify them and to compute correlations between different parameters. The statistician was blinded in terms of their identities. The ethical approval for the study was obtained from The Committee for Medical and Bioethics, Office of Research and Graduate Studies, Alfaisal University, Riyadh, Saudi Arabia.

A mini OSCE (OSCE-1) and a final OSCE (OSCE-2) were conducted where students were assessed by examiners both objectively (OSCE scores) and subjectively (ESA). OSCE-1 was a small-scale OSCE, which tested fewer but representative skills, compared with the full-scale final OSCE (OSCE-2). The OSCEs were designed by a team of expert clinical educators managing the course. A set of clinical stations were designed where the students were required to undertake a task on a standardised patient or mannequin in rotation within highly competitive time duration of a few minutes. None of the stations in OSCE-1 was repeated in OSCE-2. The set of examiners were also different in both OSCEs. However, all examiners were experienced health professionals and familiar with our OSCEs. A meeting of examiners was held prior to each OSCE to standardise the evaluation. Objective assessment was based on structured and standardised clinical checklists (see online supplementary appendix 1). The OSCE scores constituted the objective assessment, whereas an additional 'global' subjective rating (referred to as ESA) was done by the examiner for overall confidence and well-preparedness. The examiner could give 0–5 in each of the two domains, reflecting increasing expertise of the examinee. The global score from different domains was averaged out as the total ESA. Because of the flow of the students and time constraints, examiners had no opportunity to revise the scores once awarded, thus making it almost a 'first impression' grading.

Overall, each student had simultaneous SA, PA, as well as ESA at two instances, each one of which was averaged out during analysis, as shown in figure 1.

The statistical analysis was conducted using IBM SPSS V.20.0. Frequencies were calculated where relevant. The

data were checked for normality. Cronbach's alpha was calculated to check internal consistency of the subjective evaluation tool. Pearson's correlations were used to measure relationship among various parameters. The linear regression analysis was carried out to assess whether subjective assessments were predictive of the objective evaluations. Additionally, paired sample t-test was used to examine performance progression. In all analyses, only a p value <0.05 was considered significant.

RESULTS

All 164 year 2 medical students participated. There were 93 females and 71 males (57%:43%) with their ages ranging between 18 and 22 years, 55% of which were Saudis and the remaining 45% other nationalities. Their mean scores regarding various forms of assessments are given in table 1. Cronbach's alpha values for the subjective assessment tools showed acceptable reliability of SA (0.78), PA (0.87) and ESA (0.64).

Correlations

We used Pearson's correlation because the data were continuous data with no outliers.

General correlations

The scores of final comprehensive OSCE (OSCE-2) correlated positively with mini OSCE (OSCE-1) ($r=0.34$, $p<0.001$) as well as ESA ($r=0.53$, $p<0.001$). Similarly, OSCE-1 scores correlated positively with ESA ($r=0.40$, $p<0.001$).

Although SA and PA correlated to each other ($r=0.20$, $p=0.01$), there was no correlation with any OSCE or ESA.

Specific correlations

See table 2 and figure 2.

Final OSCE and mini OSCE versus ESA

The OSCE-1 is correlated with individual components of ESA, that is, self-confidence and well-preparedness

Table 1 Descriptive characteristics of students' data: means of students among all forms of used assessments

Assessment	Mean	SD
Self-assessment, average*	4.73	0.41
Self-assessment 1*	4.73	0.33
Self-assessment 2*	4.72	0.70
Peer assessment, average (maximum 5)	4.82	0.26
Peer assessment 1*	4.79	0.25
Peer assessment 2*	4.84	0.43
ESA [†]	3.92	0.43
OSCE-1 (% score)	67.94	12.76
OSCE-2 (% score)	88.64	6.59

The maximum score was 5.

[†]The ESA scores at the two instances were averaged as an overall ESA for meaningful analysis.

ESA, Examiners' Subjective Assessment; OSCE, Objective Structured Clinical Examinations.

Table 2 Correlations* between the components of self-assessment (SA), peer assessment (PA) and Examiners' Subjective Assessment (ESA)

Domain	SA versus PA	SA versus ESA	PA versus ESA
Confidence	$r=0.19, p=0.017$	$r=0.21, p=0.089$	$r=0.18, p=0.11$
Respectful manner	$r=0.19, p=0.014$		
Attentive listening	$r=0.43, p<0.001$		
No nervousness	$r=0.29, p<0.001$		
Using non-technical language	$r=0.59, p<0.001$		
Being concise (to the point)	$r=0.45, p<0.001$		
Being well-prepared	$r=0.39, p<0.001$	$r=0.23, p=0.065$	$r=0.16, p=0.425$

*The SA, PA and ESA scores at the two instances were averaged as an overall ESA for meaningful analysis.

($r=0.35, p<0.001$, and $r=0.36, p<0.001$). Similarly, OSCE-2 correlated with both components of ESA, that is, self-confidence and well-preparedness ($r=0.48, p<0.001$, and $r=0.49, p<0.001$).

Final OSCE and mini OSCE versus SA or PA

Considering the matching aspects of ESA, SA and PA, the following positive correlations were observed:

1. Well-preparedness in first SA correlated with OSCE-1 scores ($r=0.186, p=0.018$).
2. Well-preparedness in first PA correlated with OSCE-1 scores ($r=0.154, p=0.049$).
3. Well-preparedness in second SA correlated with OSCE-2 scores ($r=0.192, p=0.015$).

ESA versus SA or PA

Confidence' component of ESA

Students' SA of confidence correlated with ESA in 'confidence' ($r=0.207, p=0.008$). Both SA and PA ratings of 'no nervousness' during the session correlated with ESA in 'confidence' ($r=0.210, p=0.007$, and $r=0.185$,

$p=0.018$, respectively). Similarly, SA ratings of 'well-preparedness' during the session correlated with ESA in 'confidence' ($r=0.244, p=0.002$).

'Well-preparedness' component of ESA

Similarly, 'well-preparedness' in both SA and ESA correlated each other ($r=0.234, p=0.003$). In addition, 'well-preparedness' in ESA correlated with 'no nervousness' in SA ($r=0.191, p=0.014$).

Both 'confidence' and 'well-preparedness' in ESA correlated with each other ($r=0.662, p<0.001$).

SA versus PA

Although students' SA and PA correlated with each other in the first session ($r=0.48, p<0.001$), there was no such correlation in the second session ($p=0.80$).

Interestingly, students' first SA positively correlated with their second SA ($r=0.18, p=0.021$). However, there was no correlation between peer assessments of two sessions ($p=0.054$).

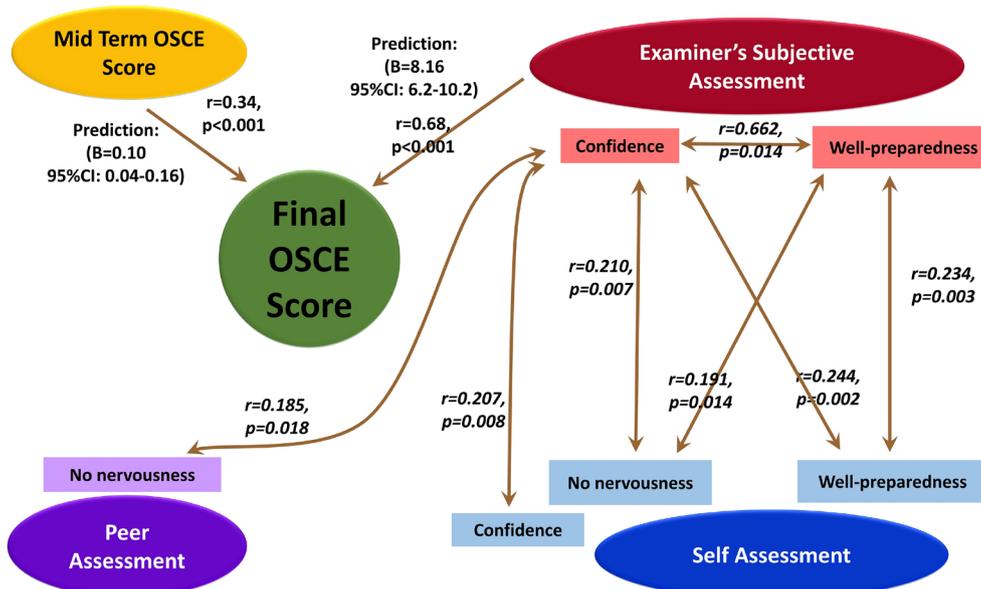


Figure 2 Predictions and correlations regarding various assessment tools and their components. The figure shows how various components of different tools relate to each other in terms of prediction (coefficient B) and correlation (Pearson's r). Neither self-assessment nor peer assessment could predict students' grades in the final Objective Structured Clinical Examinations (OSCE).

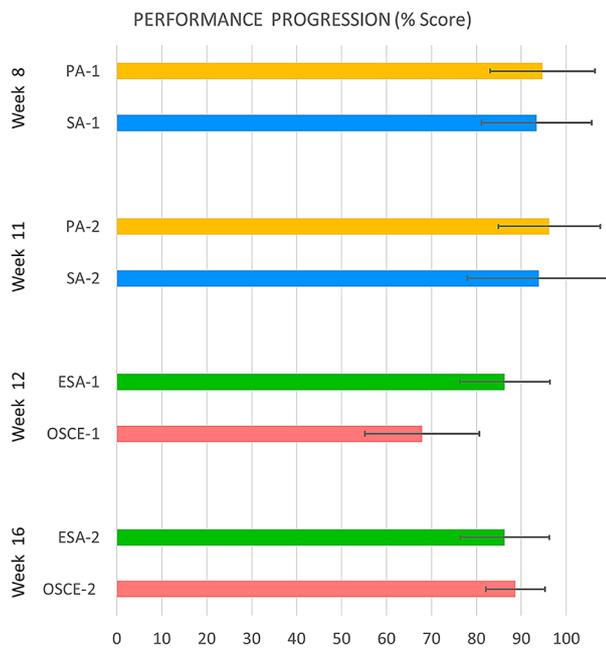


Figure 3 Performance progression of students through the course using the different assessment tools. ESA, Examiners' Subjective Assessment; OSCE, Objective Structured Clinical Examinations; PA, peer assessment; SA, self-assessment.

Performance progression

See figure 3.

We observed that there was a significant improvement in students' performance in the OSCE-2 compared with OSCE-1 ($p < 0.001$, paired sample t-test).

Both SA and PA are significantly higher than any of the subsequent ESA (p values of < 0.001 ; Wilcoxon signed-rank test).

Prediction of grades (linear regression analysis)

As shown in figure 2, ESA is a strong predictor for students' scores in the final OSCE ($p < 0.001$, $B = 8.16$, 95% CI 6.15 to 10.17).

The OSCE-1 also predicted students' performance in OSCE-2 ($p < 0.001$, $B = 0.17$, 95% CI 0.10 to 0.25). However, neither SA nor PA could predict students' scores in OSCE-1 ($p = 0.93$, $p = 0.82$) or OSCE-2 ($p = 0.39$, $p = 0.77$).

DISCUSSION

We have shown that subjective assessment, especially by students, has limited value to predict their performance in a high-stakes terminal evaluation, although it could still be used as a useful method of continuous assessment by the faculty.

Subjective tools like SA, PA and ESA have different immediate and long-term academic values^{6-11 13} despite their debated reliability.^{6 14 16 17} Hence, they are still used as student development tools.^{16 18 19} Different people may qualify a given performance by a student differently, depending on the evaluator's background, academic

level and experience.^{20 21} Yet, if the medical educators are using them, we need to understand how 'objectively' we could use such tools.

The objective evaluation of clinical skills is carried out by OSCEs which require considerable resources and cost. Since we offer clinical skills courses to medical students as junior as year 1, we also need to conduct OSCEs for a large number of students with time and place constraints. Currently > 800 medical students are enrolled in our institution. Thus, conducting an OSCE is a laborious and expensive task with our limited resources, forcing search of alternate but reliable methods of interim evaluations suitable for continuous assessment and feedback. In our case, the scores in small-scale OSCE-1 correlated well with OSCE-2 and predicted better eventual performance. Thus, it is possible that a small but appropriately designed OSCE-1 was a helpful strategy, especially in tight situations. Thus, using fewer resources, OSCE-1 gave an early, feasible and objective prediction of an individual student's performance level. Interestingly, ESA also showed to be a comparable independent predictor for the students' final OSCE score but this should be considered with caution. We used experienced and trained faculty with medical background to assess students subjectively using a simple assessment tool. In this study, 'confidence' and 'well-preparedness' were subjective domains that assessed students' global performance. This is in contrast to the reported weak correlations when subjective assessment of knowledge was compared with objective exams.²² On the other hand, Read *et al*²³ reported that such subjective tools could be reliable in experienced hands. They used checklists and global rating scales by novice as well as expert veterinarians and found out that experts assessed more reliably than novices in both objective and subjective evaluations. Another study²⁴ conducted on surgery residents also suggested that global rating scales used by experienced examiners are very reliable. In our case, the components in ESA correlated with each other, suggesting a reliable internal structure, thus making it a simple yet valuable assessment method at least for an interim analysis and feedback. However, its full utility still needs to be verified.

Like others have reported,²⁵ SA and PA correlated with each other generally as well as at the level of their subdomains in our study also. Interestingly, both SA and PA ratings are much higher than ESA. This might be due to (1) similarity of students while evaluating self or peers or (2) inflated rating of themselves or peers, as reported in the literature.^{16 18} One could argue that learners would gain more knowledge as the course continues and this could result in better correlations if conducted later; however, it should be noted that in our case the first SA and PA was carried out when about 60% of the course was completed (figure 1). In contrast, at an early stage of the course, a global rating score might not reflect students' knowledge but rather their stress or anxiety, thus potentially drawing wrong conclusions. Similarly, the second SA and PA were conducted when 75% and 80% of the course

was completed. Thus, we are confident that the timing of SA and PA was the best bet in our case. Further, we did not want the students to be biased on the basis of their OSCE results, hence such conduction of SA and PA remained most feasible approach in our case.

Among all the subjective approaches considered in our study, only ESA correlated well with the objective evaluations. In other words, ESA appears to bear an 'objective' element in it. Considering the subjectivity of SA and PA as well as lack of its correlation with ESA or OSCEs, we explored their components to develop a more reliable and somewhat 'objective' tool. Interestingly, only certain components of SA and PA seem to correlate with components of ESA (figure 2). In PA, 'no nervousness' correlated with 'confidence' in ESA, whereas in SA, 'no nervousness', 'confidence' and 'well-preparedness' positively correlate with ESA's 'confidence'. Additionally, 'no nervousness' and 'well-preparedness' in SA also correlated positively with 'well-preparedness' in ESA. This suggests that instead of complicated SA and PA tools, simpler and concise tools would be practical, reflecting better 'objectivity' in subjectivity—in this case, 'no nervousness', 'confidence' and 'well-preparedness'. Further, it could be easily used for assessment and feedback. While previous literature showed varying degrees of comparability between faculty-based versus student-based assessments,^{16 26} we propose an explanation of why in this study ESA appears to be more objective compared with SA and PA, even though they partly share similar structure and approach. The participating faculty was experienced and trained in subjective assessment; hence, these factors could enable better 'objectiveness' in their subjective assessment. This is in agreement with a study in the context of clinical clerkship.²⁷ Additionally, compared with students evaluating themselves or their peers, faculty are expected to have relatively less bias. Training to use subjective assessment and standardising definitions for each assessment domain should not be limited to faculty. Hence, if SA or PA is planned, students should also receive adequate training and preparation to use these tools. Several processes have been suggested to do this in different areas of education.^{16 28} Despite this, the ESA results need to be appreciated cautiously due to the following reason. The examiners in OSCE-2 also rated the students using a global rating scale. Being unblinded could have introduced a bias in ESA. However, it should be appreciated that the examiners were from the faculty who did not teach those students in that course; hence, their impression could well reflect on students' performance. The flow and timing of the OSCE stations were under constant check allowing little if any opportunity for the examiners to review the students' performance after they graded them, thus partly, if not fully, compensating for such a bias.

The study has some limitations. First, it is a single-institution study. Second, the objective and subjective evaluations on a given time were done by the same assessor, which could be a source of bias and could inflate

the ESA correlation with OSCE-2. Third, it was conducted on junior medical students learning the basic skills and may not be representative of more mature learners, such as residents. Fourth, the correlations are small. Although p values are significant at many places in the results, it is difficult to infer due to small correlations. However, due to the nature of the data one could come up with small correlations.²⁹ Further, our aim was to decipher whether there exists a relationship (suggested by significant p-values in our data) between a subjective and objective assessment rather than robust correlation. Future research could help fully understand such relationship. Fifth, the examiners in both OSCEs were different individuals; however, such limitation was minimised with examiner standardisation.

Overall, a carefully designed mini OSCE is not only feasible but also predictive of students' terminal evaluation. Subjective tools, despite their limited predictive value of a high-stakes examination, appear to be feasible in assessing students and providing feedback to them, at least within the context of learning basic clinical skills. ESA is one simple approach, with some degree of objectiveness. Likewise, SA and PA, being highly inflated and subjective, require continuous development and training of students. One could still argue that SA and PA is important for self-reflection as a physician while doing their clinical practice. Therefore, a combination of these tools is advised to reach sufficient objectivity, using available resources efficiently, while involving all stakeholders in learning experience and hence allowing better continuous assessment. Continuity is important because psychomotor skills and attitude groom over time. This continuity can be included into students' portfolios; in a similar manner to 'multisource feedback' discussed by others.²⁵ This study incites fresh excitement in an old discussion. Simplified subjective tools need repeated adaptation, validation, reliability check and evaluation based on the needs and settings of a course, and students' level. Training the students on using SA and PA tools can help to overcome inflated scoring and minimise bias. Using experienced faculty to use such faculty-based subjective tools is appropriate and gives reliable results. Only then we expect that these subjective tools would provide better 'objective' assessment that can be used in student grading, continuous evaluation, reflective feedback and development.

Contributors ATI, LAA and MAS participated in data analysis and manuscript write-up. AN participated in manuscript write-up. HMA, AAA and ASO participated in collection of data and drafting of the manuscript. MZ participated in drafting and organisation of the manuscript and critical appraisal of the methodology. NAA was the leader of the team that undertook this project. He contributed to conception of the idea, collection and analysis of data, drafting and organisation of the manuscript, and direct supervision of the study. He also undertook all subsequent revisions. He was also the director of the clinical skills course which was the focus of this study.

Competing interests None declared.

Ethics approval The Committee for Medical and Bioethics, Office of Research and Graduate Studies, Alfaisal University, Riyadh.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional unpublished data.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2017. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

REFERENCES

- Moore DE, Cordes DL. Needs assessment. In: Rosof AB, Felch WC, eds. *Continuing medical education - A primer*. 2nd ed. Westport, CT: Praeger, 1992:42–51.
- Tabish SA. Assessment methods in medical education. *Int J Health Sci* 2008;2:3–7.
- Gordon MJ. Self-assessment programs and their implications for health professions training. *Acad Med* 1992;67:672–9.
- Spencer JA, Jordan RK. Learner centred approaches in medical education. *BMJ* 1999;318:1280–3.
- Pisklakov S. Role of self-evaluation and self-assessment in medical student and resident education. *Br J Edu Soc Behav Sci* 2014;4:1–9.
- Speyer R, Pilz W, Van Der Kruis J, et al. Reliability and validity of student peer assessment in medical education: a systematic review. *Med Teach* 2011;33:e572–e585.
- Boud D, Cohen R, Sampson J, et al. Peer learning and assessment. *Assess Eval High Edu* 1999;24:413–26.
- Garcia G, Pearson P. Assessment and Diversity. In: Darling-Hammond L, ed. *Review of Research in education*; 20:337–9. Washington DC: American Educational Research Association, 1994.
- Pajares F. Self-Efficacy beliefs in academic settings. *Rev Educ Res* 1996;66:543–78.
- Stipek D, Roberts T, Sanborn M. Preschool-age children's performance expectations for themselves and another child as a function of the incentive value of success and the salience of past performance. *Child Development* 1984;59:1983–9.
- Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med* 1991;66:762–9.
- Boud D, Falchikov N. *Rethinking assessment in higher education*. 132. London: Kogan Page, 2007.
- Gielen S. Peer assessment as a tool for learning [dissertation]. Leuven, Belgium: Catholic University of Leuven, 2007.
- Papinczak T, Young L, Groves M, et al. An analysis of peer, self, and tutor assessment in problem-based learning tutorials. *Med Teach* 2007;29:e122–e132.
- Malau-Aduli BS, Mulcahy S, Warnecke E, et al. Inter-Rater reliability: comparison of checklist and global scoring for OSCEs. *Creat Educ* 2012;03:937–42.
- Leach L. Optional self-assessment: some tensions and dilemmas. *Assess Eval High Edu* 2012;37:137–47.
- Baxter P, Norman G. Self-assessment or self deception? A lack of association between nursing students' self-assessment and performance. *J Adv Nurs* 2011;67:2406–13.
- Strong B, Davis M, Hawks V. Self-Grading in large general education classes: a case study. *College Teaching* 2004;52:52–7.
- Mdn L, Alwis WAM, Schmidt HG. Accuracy of students' self-assessment and their beliefs about its utility. *Assess Eval High Edu* 2010;35:135–56.
- Hulsman RL, Peters JF, Fabriek M. Peer-assessment of medical communication skills: the impact of students' personality, academic and social reputation on behavioural assessment. *Patient Educ Couns* 2013;92:346–54.
- Langan AM, Shuker DM, Cullen WR, et al. Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assess Eval High Edu* 2008;33:179–90.
- Farrell TM, Kohn GP, Owen SM, et al. Low correlation between subjective and objective measures of knowledge on surgery clerkships. *J Am Coll Surg* 2010;210:680–3.
- Read EK, Bell C, Rhind S, et al. The use of global rating scales for OSCEs in veterinary medicine. *PLoS One* 2015;10:e0121000.
- Regehr G, MacRae H, Reznick RK, et al. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;73:993–7.
- Lee KL, Tsai SL, Chiu YT, et al. Can student self-ratings be compared with peer ratings? A study of measurement invariance of multisource feedback. *Adv Health Sci Educ Theory Pract* 2016;21:401–13.
- Lundquist LM, Shogbon AO, Momary KM, et al. A comparison of students' self-assessments with faculty evaluations of their communication skills. *Am J Pharm Educ* 2013;77:72.
- Goldstein SD, Lindeman B, Colbert-Getz J, et al. Faculty and resident evaluations of medical students on a surgery clerkship correlate poorly with standardized exam scores. *Am J Surg* 2014;207:231–5.
- Ross J. The reliability, validity, and utility of self-assessment. *PARE* 2006;11. <http://pareonline.net/getvn.asp?v=11&n=10>
- Murray J. Likert Data: what to use, Parametric or Non-Parametric? *JBSS* 2013;4:258–64.