BRIEF

Pass-Fail Decisions for Borderline Performers After a Summative Objective Structured Clinical Examination

Mayar Ali, BSc (Pharm),^a Shane A. Pawluk, PharmD, BSc (Pharm),^a Daniel C. Rainkie, PharmD, BSc (Pharm),^a Kyle John Wilby, PharmD, BSP^{a,b}

^a College of Pharmacy, Qatar University, Doha, Qatar

^b School of Pharmacy, University of Otago, New Zealand

Submitted October 16, 2017; accepted February 17, 2018; published March 2019.

Objective. To determine what expert assessors value when making pass-fail decisions regarding pharmacy students based on summative data from objective structured clinical examinations (OSCE), and to determine the reliability of these judgments between multiple assessors.

Methods. All assessment data from 10 exit-from-degree OSCE stations for seven borderline pharmacy students (determined by standard setting methods) and one control was given to three of eight assessors for review. Assessors determined an overall pass-fail decision based on their perception of graduate competency. Assessors were interviewed to determine their decision-making rationale. Intraclass correlation coefficients were used to calculate reliability between assessor judgments.

Results. Expert consensus was achieved for three of the eight students, however, the assessors' decisions did not align with standard-setting results. The reliability of assessors' decisions was poor. Assessors focused on ability to make correct recommendations rather than on gathering information or providing follow-up advice. Global evaluations (including a student's communication skills) rarely influenced the assessors' decision-making.

Conclusion. When faced with making pass-fail decisions for borderline students, the assessors focus on evaluating the same competencies in the students but differed in their expected performance levels of these competencies. Pass-fail decisions are primarily based on task-focused components instead of global components (eg, communication skills), despite that global components are weighted the same for scoring purposes.

Keywords: OSCE, assessment, borderline, evaluation, performance

INTRODUCTION

Objective structured clinical examinations (OSCEs) continue to be used throughout the health professions to assess whether a student is able to meet predefined clinical competencies.¹ These examinations are used in both formative and summative contexts as well as to aid in high-stakes decision-making, such as in exit-from-degree or licensure decisions.² In licensure decisions, OSCEs are used to demonstrate competence in health professional applicants and to determine which are ready to enter practice. As such, a passing score on an OSCE gives assurance to regulatory bodies and the general public regarding a student's clinical knowledge and skills. Therefore, pass-fail

criteria for these examinations must be able to accurately predict whether the candidate is adequately prepared to engage in patient care activities.³

While the depth and breadth of research on OSCEs continues to expand, there is still a gap in the literature regarding pass-fail decisions for candidates whose performance is borderline. For high-stakes assessments, these decisions must be credible as students may challenge an OSCE outcome at a higher institutional or external (legal) level.⁴ This could result in review of OSCE data by individuals other than the assessors.⁵ Thus, determining what expert assessors focus on and how they synthesize data to formulate pass-fail decisions is important. For example, an externally appointed assessor may challenge a predefined cut score for competence, depending on how he or she interprets performance in line with a program's competency expectations. This also may hold true for internally appointed assessors attempting to reconcile perceived competency discrepancies for borderline performers.

Corresponding Author: Kyle John Wilby, PO Box 56, School of Pharmacy, University of Otago, Dunedin, New Zealand. Tel: +64-3-479-7325. E-mail: kyle.wilby@otago.ac.nz

Note: At the time of manuscript submission, Dr. Wilby was affiliated with the College of Pharmacy, Qatar University.

These assessors can be thought of as “expert witnesses” who must review all evidence in order to provide an overall judgment regarding student performance. As such, it should be determined what these “expert witnesses” value in evaluation data and whether they are consistent in their ratings, values, and interpretations.

Basing pass-fail decisions solely on robust standard-setting methods will not negate the potential problems described above. The literature shows standard setting is prone to error, especially for smaller sample sizes.⁶ This means that borderline students could be inappropriately passed or failed depending on the type of error associated with the standard-setting method. Therefore, borderline performance scores may need to be reviewed on a case-by-case basis in order to arrive at more authentic performance decisions. When assessors’ decisions in borderline performance situations can be deemed credible and reliable, the assessors will have greater confidence in conducting their assessments. Understanding how assessors process data and make these judgments will address the knowledge gap of where to draw the line for borderline performers. This study sought to determine what expert assessors value when assigning pass-fail decisions based on summative OSCE data and how reliable judgments between multiple assessors are.

METHODS

This was a case series approach to determination of pass-fail decisions and assessor cognition regarding such decisions for borderline cases obtained from a summative OSCE for graduating pharmacy students. The study was conducted in Doha, Qatar, at the College of Pharmacy at Qatar University in 2016. The program enrolls 25 female students per year.

The OSCE was a third iteration of a summative high-stakes exit-from-degree examination blueprinted to examine competencies defined by the program’s competency framework, which consisted of nine interactive stations and one static station.⁷ The OSCE functioned as the final examination for the students’ final clinical course prior to graduation. Cases were blueprinted according to the program’s educational outcomes and local practice considerations. Ten case groups consisting of three to six faculty members and/or clinicians were tasked to write a case and then subsequently validate a different case written by another group. Validation occurred through role play, content review, and accuracy checking. In advance of the OSCE day, assessors and standardized actors were recruited and trained by the chief examiner. Examination center staff members, including runners, hall monitors, track coordinators, and registration personnel were also recruited. The OSCE was conducted on one

day, one track, but two cycles (a morning group of students and an afternoon group of students). Students were secluded between cycles to reduce the risk of sharing of examination materials. All examination materials and references were printed in hardcopy, and students were identified on examination booklets using coded stickers. Development procedures and evaluation results for the first and second iterations of the OSCE, which were largely similar to this iteration, have been reported previously.^{8,9}

Twenty-five graduating students from the bachelor of science in pharmacy degree completed the examination in May 2016. For this iteration of the examination, standards were set using three different methods: holistic scoring method (60%), modified Angoff method (61%), and borderline regression method (58%). Angoff standards for all cases were set by a group of five faculty members with clinical practice duties using a single focus group approach. For the borderline regression method, only data from this current iteration were included for determining standards. The borderline regression method was preplanned as the standard setting procedure for grading purposes. Upon completion of the OSCE, seven of the 25 cases were deemed borderline based on comparisons between total scores and established cut scores. Of these seven cases, all students passed the examination according to the borderline regression method, five students failed according to the holistic method, and seven students failed according to the modified Angoff method. As no clear pass-fail decision could be made for these students, the seven students in question were deemed “borderline.”

For the present analysis, we selected the seven cases that were determined borderline, as well as one control case in which the student clearly passed according to all methods. The evaluation data for these eight cases were extracted from the overall OSCE dataset. Data included task-oriented checklists and global assessment rubrics for each of the nine interactive stations but did not include video or performance observations. The checklist for each station consisted of 8 to 20 points (eg, “asked about allergies” or “provided accurate dosing instructions”) the student must say or do during the allocated time. Points were distributed between sections of gathering information, recommendations/management, and monitoring/follow up. Students either received a full point or no point based on their assessed performance. No partial points were given. The global assessment (representative of communication skills) consisted of a five-point rubric and a comment field asking evaluators to justify their ranking. The checklist and global assessment were each weighted 50% of the total score for each station. One static written station was included that required students to check the

accuracy of a calculated dose for a filled prescription. Students received points for identifying the error, justifying the existence of the error, and correcting the error. This was completed on paper and the student answers and feedback were included in the evaluation packages extracted for the present study.

Eight expert clinical assessors were recruited to determine overall pass-fail decisions for the eight selected student cases, and assessors' written consent was obtained. This number was chosen to ensure each student was assessed by three independent assessors, which is in line with previous similar studies.¹⁰ Assessors were purposively sampled based on previous experience evaluating students during OSCEs and clinical training settings and familiarity with the competency expectations of students upon graduation. Assessors also were chosen based on differing clinical practice area to promote diversity within the assessor group. Assessors were eligible for participation if they were registered pharmacists and if they were affiliated as faculty members or clinical adjunct faculty members with the College of Pharmacy at Qatar University. Assessors received training during a 15-minute face-to-face meeting with study investigators on an individual basis. Each assessor received the examination blueprint and all evaluation data for three students (blinded), based on a distribution matrix developed that ensured no assessors evaluated the same three students. As data were limited to checklists and rubric scores, the chance of bias from unblinding was negligible. Assessors were instructed to review all evaluation data available and to determine if they believed the student should pass or fail the OSCE as a whole. They were also informed that they would be interviewed about their rationale for decision-making and therefore could keep notes about why they made certain decisions if needed. Once pass-fail decisions were returned, data were input into a database using Microsoft Excel. Interrater reliability for pass-fail decisions was calculated using intraclass correlation coefficients (two-way random model) using IBM SPSS statistics, version 24 (IBM, 2016).

Within three days of receiving an assessor's pass-fail decisions, he or she was contacted for a semi-structured interview to explore their reasoning for their overall judgments. Two investigators conducted the interviews according to a predefined script. Each interview lasted approximately 15 minutes. Interviews were recorded and intelligent transcripts were produced immediately afterwards. A second investigator reviewed each transcript for transcription errors. Two investigators independently coded each transcript using a bottom-up, open coding approach. Coding was reviewed after each transcript to discuss code labels and rectify any coding discrepancies. A third investigator was available to resolve

any discrepancies that could not be resolved between the two coding investigators, but the need for this investigator did not occur. Once all transcripts were coded, codes were combined into related categories and data were reviewed and interpreted for themes. All investigators had access to the data and all investigators agreed upon final included themes.

The study was exempted from full review by the Qatar University Review Board.

RESULTS

Pass-fail results per assessor and overall reliability ratings are provided in Table 1. Three students achieved 100% consensus regarding the pass-fail decisions between the three assessors (two pass and one fail decision). One of these students was the control student who was expected to pass. The remaining five students did not reach complete agreement among the three assessors, with four students (80%) receiving two pass decisions and one fail decision and one student (20%) receiving one pass decision and two fail decisions. The intra-class correlation coefficient was 0.407, translating to suboptimal reliability. The following major themes were interpreted from the data pertaining to assessors' rationale for making pass-fail decisions: assessors value quantitative data when making evaluation decisions; assessors value what students do more than how they do it; and assessors value key checklist recommendation and management strategies. Each theme is further discussed below and supporting quotations are provided in Appendix 1.

The first theme identified was that assessors value quantitative data when making evaluation decisions. Quantitative strategies were used overwhelmingly by the assessors in making an overall pass-fail decision. Strategies included counting the number of stations a student passed, counting the number of checkmarks achieved by students for each station, and setting thresholds for the number of

Table 1. Pass-Fail Designations by Three Assessors for Each Borderline Student

Student	Passing Designations	Failing Designations
1	1	2
2	2	1
3	0	3
4	3	0
5 ^a	3	0
6	2	1
7	2	1
8	2	1

Abbreviations: F=Fail; P=Pass.

^a Control student.

stations a student should be allowed to fail yet still pass the examination. Assessors focused on the number of stations a student was successful at completing versus the content of the stations themselves. Assessors did not always formulate decisions using the same methods. For example, assessors who counted the number of stations a student passed may not have set different thresholds for the number they allowed the students to fail before assigning an overall fail rating for the OSCE.

The second theme identified was that assessors value what students do over how they do it. Assessors highly valued the content portion of the evaluation package. The global assessment was typically consulted only if content-oriented analytical checklists were not conclusive of performance. When asked specifically regarding the global assessment, many assessors stated having difficulty with how to interpret an overall performance rating focused on communication versus a skills-based checklist of what students did or did not do. Again, assessors tended to count checkmarks in order to determine if a student passed or failed a particular station.

The third theme identified was that assessors based their overall decision-making on checklist points relating to recommendations or management strategies. Assessors overwhelmingly valued the recommendation/management section of the analytical checklist when formulating their overall pass-fail decisions. Assessors focused on these sections when reviewing stations and used these sections to decide on a pass-fail decision per station and overall. Other sections, such as monitoring/follow up and gathering information were not very influential to the assessors' overall decisions.

DISCUSSION

The purpose of this study was to determine how reliable assessors are at determining pass-fail decisions for borderline students after completion of a summative OSCE and to evaluate how they arrive at such decisions after reviewing assessment data. Results showed that assessors often came to different conclusions when making decisions regarding students whose performance was borderline; however, assessors' decisions were more consistent when evaluating a control student who performed well. Assessors largely based their decisions on quantitative measures of performance, including counting checkmarks and the number of stations the student completed accurately. Assessors did not place high value on more global assessments. Finally, assessors greatly valued the recommendation/management component of the checklist evaluation and did not value how students gathered information or provided monitoring/follow up recommendations.

The major finding of this study was that assessors arrive at pass-fail judgments differently, yet focus on the same aspects of performance when reviewing evaluation material from a summative OSCE. The rationale for this finding could be explained by a number of factors, including varying competency expectations among assessors, differing thresholds for summative decision making (ie, number of failed stations they allowed), the nature of evaluating a truly "borderline" student, or assessor error. Based on the discrepancies observed using multiple standard-setting techniques, in addition to the poor reliability found in this study, these students' performance likely was truly borderline and thus challenging to assess. The implications of this finding are broad, as they suggest that expert review of evaluation material may not suffice in terms of being defensible. If this examination was truly a "must pass" examination for graduation, it would be most appropriate to act in the best interest of the students and pass all students in this study, especially as they all passed according to the preplanned borderline regression method. Otherwise, there is not enough evidence to deem any of these students' performance a "failure." However, if a student had failed the examination according to all accepted standard-setting methods, assigning a failing grade would be justified. For the purposes of the OSCE in this study, all students were deemed to pass as the borderline regression method was pre-selected as the grading mechanism.

The second major finding from this study was that assessors almost entirely focused on "recommendation/management" points of the task-completion checklist to determine overall competency for each station. Other sections of gathering information, monitoring and follow up were largely ignored. Additionally, global performance ratings were only referred to in a small number of cases. This finding suggests that assessors, in a summative context, overwhelmingly value what a student does or does not do more than they value how the student does it or if the student is perceived to do it effectively. This finding suggests that assessors deem competence based on outcome rather than process. This result has significant implications. For example, a student could be deemed competent because she made a correct recommendation (ie, to provide the patient with a referral) when in fact the student's communication skills were ineffective and, as a result, in an actual practice setting a patient might not follow her advice. This could lead to patient harm, yet as a student the pharmacist was deemed "fit to graduate." As such, expert review of assessment data may need to include weighting components that require assessors to judge all aspects of student performance. These results may also support the notion that assessors who actually

witness the student interaction should provide a more holistic evaluation that can better inform whether the student completed the case successfully, considering all required competencies.¹¹ Based on this finding, this type of assessment will be included as evaluation data in future iterations of the OSCE.

Our results align with those found in the growing body of literature relating to assessor reliability and rater cognition that has identified concerns regarding accuracy, validity, and reliability of student evaluations.¹² Researchers have found that assessors focus on different aspects of performance and arrive at their own unique interpretations and judgments, even when watching the same interaction or task.^{13,14} Although our findings are specific to post hoc evaluation of assessment data, we too found that assessors interpret student performance data differently. The interesting caveat of our study, however, is that assessors primarily focused on the same evaluation components when arriving at their pass-fail decisions. Therefore, assessors in our context likely have differing expectations of what it means to be a competent pharmacist upon graduation of an entry-to-practice degree in Qatar. Previous reports found that attempts to standardize these expectations and judgments have limited success, which supports our previous suggestion that repeated measurements in an attempt to saturate assessment data may allow for greater reliability in pass-fail judgments.¹⁵

Findings of this study must be interpreted in light of some limitations. First, our sample size was small and may not be representative of all borderline cases for all graduated students, but the pervasive inconsistency we found in assessor judgments supports our findings that expert determination of pass-fail decisions is not reliable based on review of evaluation data alone. For future studies, we recommend including more control cases, in order to make cross-group comparisons and better understand the influence of borderline scores on assessor discrepancies. Second, this was the first time assessors were asked to make overall OSCE judgments, and some assessors, as mentioned before, may be biased towards passing borderline students because of the summative nature of the OSCE in this context. Nevertheless, these limitations do not preclude us from formulating meaningful conclusions based on the results obtained.

CONCLUSION

The results of this study provide greater understanding about how assessors synthesize summative OSCE data for borderline students and how they arrive at performance judgments. Although assessors used different

decision-making processes when reviewing assessment data, they focused on the same evaluation components. This discrepancy in competency judgments has negative implications for pass-fail decision-making. These results should encourage programs to develop firm competency expectations, with subsequent training of assessors on how to measure student performance both while observing interactions and evaluating students post hoc. Assessors also focus on what was done rather than how it was done or if it was done effectively. Therefore, assessment tools should be refined to better capture an overall measure of performance in order to avoid “checkmark counting” by assessors and potentially rewarding behavior that might actually lead to patient harm in an actual patient encounter.

ACKNOWLEDGMENTS

Funding was provided by a grant from Qatar University.

REFERENCES

1. Newbie D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ.* 2004;38(2): 199-203.
2. Austin Z, O’Byrne C, Pugsley J, Munoz LQ. Development and validation processes for an objective structured clinical examination (OSCE) for entry-to-practice certification in pharmacy: the Canadian experience. *Am J Pharm Educ.* 2003;67(3):Article 76.
3. Terry R, Hing W, Orr R, Milne N. Do coursework summative assessments predict clinical performance? A systematic review. *BMC Med Educ.* 2017;17:40.
4. Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med.* 2000;75(3):267-271.
5. Tweed MJ, Thompson-Fawcett M, Wilkinson TJ. Decision-making bias in assessment: the effect of aggregating objective information and anecdote. *Med Teach.* 2013;35(10):832-837.
6. Homer M, Pell G, Fuller R, Patterson J. Quantifying error in OSCE standard setting for varying cohort sizes: a resampling approach to measuring assessment quality. *Med Teach.* 2016;38(2): 181-188.
7. Association of Faculties of Pharmacy of Canada. Educational outcomes. <http://afpc.info/node/39>. Accessed October 12, 2017.
8. Wilby KJ, Black EK, Austin Z, Mukhalalati B, Aboulsoud S, Khalifa SI. Objective structured clinical examination for pharmacy students in Qatar: cultural and contextual barriers to assessment. *East Mediterr Health J.* 2016;22(4):251-257.
9. Sobh AH, Austin Z, Izham MI, Diab MI, Wilby KJ. Application of a systematic approach to evaluating psychometric properties of a cumulative exit-from-degree objective structured clinical examination (OSCE). *Curr Pharm Teach Learn.* 2017;9(6):1091-1098.
10. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ.* 2015;49(3):296-306.
11. Jefferies A, Simmons B, Tabak D, et al. Using an objective structured clinical examination (OSCE) to assess multiple physician competencies in postgraduate training. *Med Teach.* 2007;29(2-3): 183-191.

12. Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ Theory Pract.* 2007;12:239-260.
13. Wilby KJ, Govaerts MJB, Austin Z, Dolmans DHJM. Exploring the influence of cultural orientations on assessment of communication behaviours during patient-practitioner interactions. *BMC Med Educ.* 2017;17:61.
14. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Health Sci Educ Theory Pract.* 2013;18(3):325-341.
15. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ.* 2014;48(11):1055-1068.

Appendix 1. Supporting Quotes for Each Identified Theme

Theme 1: Assessors value quantitative data when making evaluation decisions

- "I counted the number of ticks, the number of maybes, and the number of crosses. And then I went back to the maybes, I went over them again and made the decision whether they got all the information, all the required information or whether they would fail the station." (Assessor 6)
- "I guess to quantify it in my mind, of how, what would be like a pass fail, I kind of said that, ok if they fail 3 station they would be a fail versus if it was just 1 or 2 then they would pass." (Assessor 1)

Theme 2: Assessors value what students do, as compared to how they do it

- "I focused more on the analytical checklist rather than the global, because for me just having good communication or a professional attitude with the patient doesn't necessitate that the student should pass." (Assessor 2)
- "I focused more on [analytical]. Because you can communicate very well doesn't mean you can provide the best answer or the correct answer, so I put more weight on the analytical than I did on the global assessment." (Assessor 4)

Theme 3: Checklist points relating to recommendations or management strategies comprise the primary basis for assessors' overall decision-making

- "Overall, I would look at their total analytical checklist grades. If they passed and got the recommendations/management that was a solid pass. If they got all the points but missed the main recommendation/management, I would fail them." (Assessor 2)
- "I looked for the overall purpose of the case, so in terms of they are supposed to make a recommendation on a certain drug, or supposed to refer and if they achieved that portion of it I will consider it the most." (Assessor 4)