A Comparison of Standard-setting Procedures for an OSCE in Undergraduate Medical Education

David M. Kaufman, EdD, Karen V. Mann, PhD, Arno M. M. Muijtjens, PhD, and Cees P. M. van der Vleuten, PhD

ABSTRACT

Purpose. To compare four standard-setting procedures for an objective structure clinical examination (OSCE).

Methods. A 12-station OSCE was administered to 84 students in each of the final (fourth-) year medical classes of 1996 and 1997 at Dalhousie University Faculty of Medicine. Four standard-setting procedures (Angoff, borderline, relative, and holistic) were applied to the data to establish a cutoff score for a pass/fail decision.

Results. The procedures yielded highly inconsistent results. The Angoff and borderline procedures gave similar results; however, the relative and holistic methods gave widely divergent results. The Angoff procedure yielded results reliable enough to use in decision making for a high-stakes examination, but would have required more judges or more stations.

Conclusions. The Angoff and borderline procedures provide reasonable and defensible approaches to standard setting and are practical to apply by non-psychometricians in medical schools. Further investigation of the other procedures is needed.

Acad. Med. 2000;75:267–271.

Introduced over 15 years ago,[1] objective structured clinical examinations (OSCEs) are increasingly used in U.S. and Canadian medical schools; in 1994, 111 schools reported using OSCEs and standardized patients (SPs) to assess their students.[2] SP-based, multiple-station OSCEs are now a part of several high-stakes examinations, including the Canadian qualifying examination[3] and an examination for international medical graduates wishing to practice in Canada[4]; The National Board of Medical Examiners is now considering the use of OSCEs in the U.S. licensing examination. Despite this increased use and its accompanying plethora of studies, some issues surrounding the use of OSCEs remain unanswered.[5] Particularly deserving of attention in such high-stakes examinations as those mentioned above are the procedures for standard setting. To help fill this gap in the literature, we compared various standard-setting procedures, specifically investigating which procedures would be most effective in establishing an appropriate cutoff score for a pass/fail decision in a multistation OSCE.

Many studies of OSCEs and of SPs have been reported since their inception.[4,6,7] Researchers have investigated factors that influence reproducibility, including examinees' performances across stations, inter-rater reliability, differences in SPs playing the same role, and examination and station length.[6]

Until recently, rather less attention has been devoted to standard-setting procedures for SP-based examinations, analogous to those available for written tests.[6,8] LaDuca and colleagues[9] discussed strategies for setting standards for performance assessments of physicians' clinical skills. They outlined traditional and alternative approaches to setting standards, compared score-based versus content-based standards, and showed how to apply the Angoff method (described below) to SPs.

In his comprehensive review of the literature on standard setting,[10] Cusimano asked: "Standard setting is the process of deciding 'what is good enough.' How do we actually make such a decision, when by all conceptions, competence is a continuous variable?" Cusimano referred to the standard as a conceptual boundary (on the true-score

***Dr. Kaufman** is professor, and **Dr. Mann** is professor and director, both in the Division of Medical Education, Dalhousie University, Halifax, Nova Scotia. **Dr. Muijtjens** is assistant professor, Department of Medical Informatics, and **Dr. van der Vleuten** is professor and chair, Department of Educational Research and Development, University of Maastricht, Maastricht, The Netherlands.*

Correspondence and requests for reprints should be addressed to Dr. Kaufman, Division of Medical Education, Clinical Research Centre, Rm. C-115, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7; e-mail: ⟨david.kaufman@dal.ca⟩.

scale) between acceptable and non-acceptable performances, while a passing score is a particular point (on an observed-score scale) that is used to make decisions about examinees.

Methods of standard setting have been divided into three groups: judgmental methods, empirical methods, and combination methods. Judgmental methods inspect individual test items to judge how the minimally competent person would perform on each item. Empirical methods, in contrast, require examinee test data as part of the standard-setting process. Combination methods use both empirical and judgmental data.

Judgmental procedures for standard setting include those described by Angoff,[11] Ebel,[12] Nedelsky,[13] and Jaeger.[14] In the Angoff method, judges examine each item and decide the probability that a minimally competent candidate would answer that item correctly; group discussion may follow, not necessarily leading to consensus. In some cases, a second round of judgment of each item follows the group discussion.[15] The sum of the final judgments represents the minimally acceptable score. The test standard is the average of the sums for the sample of judges. The Angoff method is easy to implement and it is easy to compute the cutoff scores; thus, this method is very popular.

Within the empirical methods, an assessment is made of examinees' performances, either as individuals or as a group. Judgments of test content are not used directly in this method. In the borderline-group method, the mean or median score of performances identified as minimally acceptable or "borderline" is used to derive cutoff scores both for individual stations and for the total test. This method requires judges to determine what they consider to be borderline performance based on their knowledge of the domains tested and of examinees' performances in those domains.

The combination method involves an empirical statistical approach, based-on a cutoff score established by one of the methods above. This method uses mathematical modeling to minimize incorrect classifications. Clauser and colleagues[16] found a simple regression model correlated better with expert ratings of the same performance than did a rule-based model when applied to computer-based examination (CBX) cases.

Studies have shown that different standard-setting methods may produce quite different results, particularly if different sets of judges are used for each method. Rothman and Cohen[17] compared empirically and rationally defined standards for clinical skills checklists. They found that the judges were essentially responding to two different tasks. The empirical standards were thought to be more acceptable in terms of pass rates and comparison with previous candidates' performances. Several studies have also examined how judgments for individual cases relate to overall test decisions.[18–20] Judges appeared to use both compensatory methods of establishing test-level decisions and mixed compensatory and non-compensatory approaches. These different approaches may result in different classifications of candidates, particularly where case specificity is involved. Based on his comprehensive review of studies of standard-setting methods in evaluating physician performance,[10] Cusimano emphasized the need for more research into these methods, particularly in the area of the OSCE.

This study addressed two research questions: (1) What is the reliability of the Angoff passing score as a function of the number of judges and stations? (2) What are the differential outcomes of several standard-setting methods for the passing score and failure rate?

Method

Participants

The participants in this study were two cohorts of 84 students each (classes of 1996 and 1997); all students were in their fourth (and final) year of a problem-based learning (PBL) undergraduate medical curriculum at Dalhousie University Faculty of Medicine in Halifax, Nova Scotia, Canada.

Procedures

A centrally organized OSCE was administered in April of 1996 and 1997. The examination included 15 stations: five ten-minute and ten five-minute stations. Three of the short stations were written stations that did not involve SPs; they were therefore excluded from this study, leaving a total of 12 stations. All students were required to take this evaluation; however, it was not a pass/fail examination and did not affect the ability to graduate. The examination was developed to follow closely the format of the Medical Council of Canada (MCC) Part II OSCE, which all candidates must pass approximately 17 months after entering graduate medical education. The stations included a selection of skills that represented the entire clerkship experience; three stations tested communication skills. The contents of the 12 stations (and each station's length in minutes) were: brain-dead husband (ten); depression (ten); elderly parent brought to emergency department by daughter (ten); chest pain (ten); acute abdominal pain (ten); heavy smoker (five); hematuria (five); annual breast exam (five); teenage boy with twisted knee (five); 16-year-old girl with cerebellar disorder (five); mother with three-year-old boy with ear pain (five); 23-year-old woman with abnormal gynecologic bleeding (five).

Clinical faculty from a variety of disciplines served as raters. Each ten-minute station was replicated four times to create four identical, concurrent tracks. Each track used a different clinician-rater for the station. Each five-minute station was duplicated twice, and one clinician–rater was needed for each of the two. A total of 34 clinician–raters

participated in the 12 stations for the morning session, in which half the class participated. Some raters were replaced in the afternoon session; approximately half stayed for both sessions. Many different examiners were used in the second (1997) administration of this OSCE. Most were experienced raters, having participated previously in the MCC Part II qualifying examination. The SPs were trained by the medical school to the level required for the MCC examination. The stations remained unchanged for 1996 and 1997. At each station, examiners completed a checklist; in addition, they provided a global rating of either "pass," "borderline," or "fail." An examinee's score for a station, used in the analysis, was the percentage score on the checklist for that station. The overall test score was defined as the mean of an examinee's 12 individual station scores.

Standard-setting Procedures

We used five standard-setting approaches: the Angoff method, the borderline method, two relative methods, and a holistic method with a pre-established passing score of 60%. Each is described below.

The Angoff method. Five raters participated in this procedure. Two were faculty members (pediatrics and emergency medicine), three were final (fourth- or fifth-) year residents (urology, surgery, and medicine). All participants were experienced with our curriculum and had taught students at the fourth-year clerkship level. The two faculty persons also had considerable experience with the OSCE format and with SPs. The group reviewed the standard-setting method to be used. Through discussion, they reached consensus on a definition of a minimally acceptable "borderline" candidate. Using that definition, each rater rated each station independently, answering the following question: "Think of a group of borderline candidates. What proportion of them will be able to successfully pass this station?" Following the individual ratings, the group gathered and displayed the ratings, then discussed the reasoning behind any discrepancies. Following the discussion, each rater again rated each station, answering the same question. All ratings were collected again.

The borderline method. In this method, previously described,[17] each examiner, in addition to completing the station checklist, provided an overall rating of "outstanding," "clear pass," "borderline," or "clear fail." The distribution of the scores of all "borderline" candidates for each station was calculated. The mean score was established as the standard for that station. The overall pass standard for the test was obtained by calculating the mean of the mean scores for all stations.

The relative methods. In the first of two relative methods, we took the mean of the score distribution for the group as a reference, then picked a point below that mean as the passing mark. We used the "Wijnen method"[21] (1.96 times the standard error of measurement below the mean) to determine that passing mark. This method takes into consideration the reliability of the examination. In an unreliable examination, the pass score will be more lenient, and students will not be victimized. The disadvantages of the method are that a fixed percentage of students fail and students may influence the passing score by deliberately scoring poorly (although this is not likely).

The second relative method took the best students as a reference point, since those students in general are well prepared for the examination and ambitious to obtain their optimum scores. Therefore, fluctuations in the scores of this group of students are assumed to reflect fluctuations in examination difficulty or curriculum quality, rather than student performances. An arbitrary minimum percentage level is defined below the reference score. We used a passing score that was 60% of the 95th percentile rank score of the group.[22]

The holistic method. Using the medical school's faculty-wide pass mark, the total score across all stations required to pass the examination was 60%.

Results

The Angoff method. The Angoff procedure yielded a mean passing score of 52.00% before discussion by the panel of judges and 51.17% after discussion. However, 14.3% of total variance in the Angoff standard was attributable to the main effect of variation among the five judges before discussion (i.e., their systematic leniency or severity across candidates). After discussion, there was zero variance in the Angoff standard due to variation among judges. Therefore, the results reported here will be based upon data obtained from the judges after discussion (the second rating). The percentage of total variation among stations was 57.9%, with 42.1% due to error variance. This indicates that a wide range of station difficulties was found in the OSCE and that there were only small differences among the Angoff estimates of the judges for the total examination. However, the error variance was relatively large, suggesting that there were probably considerable differences among judges in the Angoff estimates per station.

Table 1 shows the root-mean-square error (RMSE) of the test's passing score as a function of the number of judges and the number of stations in the OSCE. The RMSE is the error of the OSCE's passing score expressed on the original percentage scoring scale. With five judges and 12 stations, the RMSE was 1.45%. This yields a 95% confidence interval for the examination's passing score of 51.17% ± 2.90, which was rather large compared with the standard deviation of the actual examination scores of 5.34% (with a mean score of 63.21%).

The Angoff passing score (second

rating) of 51.17% resulted in a failure rate of 0.65%. Within the 95% confidence interval, the failure rate was found to vary from 0% to 5.16%. When the RMSE is reduced to 1%, the confidence interval narrows to 51.17% ± 2.0, and the failure rate varies from 0% to 3.23%. For a test consisting of 12 stations, Table 1 shows that this increase of passing score accuracy would require a panel of at least ten judges; with the current panel of five judges, at least 24 stations would be required.

The borderline method. The mean and standard deviation for all borderline students were calculated for each station and then for the complete examination. The mean value for the total test was 52.46%, and the average standard deviation was 9.74%.

The relative methods. The mean and standard deviation of the score distribution were 63.21% and 5.34%. The reliability (alpha) of the test was 0.517, so the corresponding standard error of measurement was 3.71%, yielding a passing score of 55.94% (for 1.96 SEMs below the mean) and a failure rate of 8.39%. The second relative method yielded a 95th percentile rank score of 72.27%, which leads to a passing score of 43.36% (.60 × 72.27%) and a failure rate of 0%.

The holistic method. Using the faculty-wide standard, a passing score of 60% was applied to the OSCE, resulting in a failure rate of 26.45%.

Table 1

Root-mean-square Error (RMSE) of the Angoff Standard of the Complete Test after Discussion (Second Rating), as a Function of the Numbers of Stations and Judges, on a 12-item OSCE Administered at Dalhousie University Faculty of Medicine, 1996 and 1997*

No. of Stations	Number of Judges									
	2	3	4	5	6	7	8	9	10	11
4	3.98	3.25	2.81	2.52	2.30	2.13	1.99	1.88	1.78	1.70
8	2.81	2.30	1.99	1.78	1.62	1.50	1.41	1.33	1.26	1.20
12	2.30	1.88	1.62	1.45	1.33	1.23	1.15	1.08	1.03	0.98
16	1.99	1.62	1.41	1.26	1.15	1.06	0.99	0.94	0.89	0.85
20	1.78	1.45	1.26	1.13	1.03	0.95	0.89	0.84	0.80	0.76
24	1.62	1.33	1.15	1.03	0.94	0.87	0.81	0.77	0.73	0.69
28	1.50	1.23	1.06	0.95	0.87	0.80	0.75	0.71	0.67	0.64

*In the Angoff method,[15] judges examine each item and decide the probability that a minimally competent candidate would answer that item correctly; group discussion may follow, not necessarily leading to consensus. In some cases, a second round of judgment of each item follows the group discussion. The sum of the final judgments represents the minimally acceptable score. The test standard is the average of the sums for the sample of judges.

Table 2

Standard-setting Procedures Applied to the 12-item OSCE Administered at Dalhousie University Faculty of Medicine, 1996 and 1997

Standard-setting Procedure	Passing Score (%)	No. of Failures	Failure Rate (%)
Angoff (second rating)	51.17	1	0.65
Borderline	52.46	3	1.95
Relative Wijnen	55.94	13	8.39
Relative 95th percentile	43.36	0	0.00
Holistic (absolute)	60.00	41	26.45

Discussion

Table 2 shows the diverging results we obtained with the different standard-setting procedures. The judgmental methods (Angoff and borderline) resulted in low failure rates (under 2%), whereas the holistic method resulted in a high failure rate of 26%. This might indicate, respectively, that the judges rated too leniently or that the OSCE really is too difficult to comply with an absolute passing score of 60%. The relative of Wijnen method resulted in a failure rate of 8%. This should not be too surprising, because by definition this method yields a failure rate between 2.5% (SEM equal to the SD of the scores) and 50% (SEM equal to zero) provided that the test scores are approximately normally distributed. The second relative method states than an examinee should pass when his or her result is higher than 60% of the 95th percentile level of the test score distribution. As a consequence, a relatively narrow distribution of scores is bound to result in zero failures, as is the case with this investigated OSCE.

Conclusion

This study has demonstrated that the ten-year-old comments of van der Vleuten and Swanson[6] still ring true—that procedures for setting pass/fail standards on SP-based tests remain primitive. It appears that a reasonably fair and accurate pass standard can be established using an Angoff procedure. However, a larger number of judges or stations would be required to obtain an acceptable level for the reliability of the pass/fail standard in the OSCE. The borderline method appears to give valid re-

sults, is simpler to apply, and has been used extensively. However, we do not have information regarding its reliability, as we have with the Angoff method, so we cannot compare the reliabilities of the two methods. The relative Wijnen and relative 95th percentile methods yielded extremely different results, both at odds with the Angoff and borderline methods. Therefore, more investigation needs to be done with the relative methods. Finally, the holistic method appears to be inappropriate, as it can lead to a severely high failure rate (e.g., 26% in this study) if the standard is applied too rigidly. The failure rate could be adjusted by setting a lower pass score, but this also would be arbitrary. However, its advantage is that, if a failure rate is set in advance, the pass score can be adjusted appropriately.

We conclude that the Angoff and borderline methods provide reasonable and defensible approaches to standard setting and are practical to apply by non-psychometricians in medical schools. From a cost perspective, the borderline method is to be preferred. The educational benefits of the OSCE for providing practice and feedback within medical schools make it a useful tool, despite its psychometric difficulties. These difficulties in standard setting can be easily overcome in low-stakes decisions about students. However, for high-stakes situations, e.g., licensing and recertification, further investigation of these methods is necessary.

The authors thank the Medical Council of Canada for the funding to support this study, and Nancy Ruedy for her excellent work in organizing and running the OSCE.

References

1. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination. Med Educ. 1979;13: 41–54.
2. Anderson RM, Stillman PL, Wang Y. Growing use of standardized patients in teaching and evaluation. Med Educ. 1994;5:15–22.
3. Reznick RK, Blackmore D, Cohen R, et al. An objective structured clinical examination for the licentiate: report of the Medical Council of Canada; from research to reality. Acad Med. 1992;67:487–94.
4. Vu NV, Barrows HS. Use of standardized patients in assessments: recent developments and measurement findings. Educ Res. 1994; 23–30.
5. Dauphinee WD, Blackmore DE, Smee S, Rothman AI, Reznick R. Using the judgments of physician examiners in setting the standards for a national multi-center, high-stakes OSCE. Adv Health Sci Educ. 1997;2: 201–11.
6. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med. 1990;2:58–76.
7. Swanson DB, Norman GR, Linn RL. Performance-based lessons from health professions. Educ Res. 1995; June/July:5–11, 35.
8. Norcini J, Lipner RS, Langden CA. A comparison of three variations on a standard-setting method. J Educ Meas. 1987;24: 56–64.
9. LaDuca A, Klass D, Downing SM, Norcini J. Strategies for Setting Standards on Performance Assessments of Physicians' Clinical Skills. Workshop presented at the Annual Meeting of the Generalists in Medical Education, New Orleans, LA, November 1992.
10. Cusimano M. Standard-setting in medical education. Acad Med. 1996;71(10 suppl): S112–S120.
11. Angoff WA. Scales, norms and equivalent scores. In: Thorndike RL (ed). Educational Measurement. Washington, DC: American Council on Education, 1971.
12. Ebel RL. Essentials of Educational Measurement. 3rd ed. Englewood Cliffs, NJ: Prentice–Hall, 1979.
13. Nedelsky L. Absolute grading scores for objective tests. Educ Psychol Meas. 1954;14: 13–9.
14. Jaeger RM. Certification of student competence. In: Linn RL (ed). Educational Measurement. New York: MacMillan, 1989:485–515.
15. Shephard LA. Setting performance standards. In: Berk RA (ed). A Guide to Criterion-referenced Test Construction. Baltimore, MD: Johns Hopkins University Press, 1980: 169–98.
16. Clauser BE, Ross LP, Fan YV, Clyman SG. A comparison of two approaches for modeling expert judgment in scoring a performance assessment of physicians' patient management skills. Acad Med. 1998;73(10 suppl):S117–S119.
17. Rothman A, Cohen R. A comparison of empirically- and rationally-defined standards for clinical skills checklists. Acad Med. 1996; 69(10 suppl):S1–S3.
18. Ross LP, Clauser BE, Margolis MJ, Orr NA, Klass DJ. An expert-judgment approach to setting standards for a standardized patient examination. Acad Med. 1996;71(10 suppl): S4–S6.
19. Clauser B, Orr N, Clyman SG. Models for making P/F decisions for performance assessments involving multiple cases. In: Rothman A, Cohen R (eds). Proceedings of the Sixth Ottawa Conference on Medical Education. Toronto, ON: University of Toronto, 1995: 239–42.
20. Margolis M, de Champlain AF, Klass DJ. Setting examination-level standards for a performance-based assessment of physicians' clinical skills. Acad Med. 1998;73(■ suppl):S114–S116.
21. Wijnen WHFW. Onder of Boven de Maat [Missing or Hitting the Mark]. PhD dissertation, University of Groningen, Groningen, The Netherlands, 1971.
22. Cohen-Schotanus J, van der Vleuten CPM, Bender W. Een betere cesuur bij tentamens: de beste studenten als referentiepunt [A better cutoff for examinations: the best students as a reference point]. In: Ten Cate TJ, Dikkers JH, Houtkoop E, Pollemans MC, Pols J, Smal JA (eds). Gezond Onderwijs [Health Education], vol. 5. Houten/Diegem, The Netherlands: Bohn Stafleu van Loghum, 1996:83–8.