Enhancing the defensibility of cut scores: A multi-source evaluation of a modified Angoff standard-setting process in a high-stakes medical examination

Abstract

Introduction: Standard setting is essential in health professions education, yet establishing credible and defensible cut-off scores remains challenging, particularly when post-standard-setting evaluation is limited.
Objective: To apply a systematic, multi-source approach to enhance the validity and defensibility of cut-off scores derived through a modified Angoff procedure in a high-stakes medical examination.
Methods: A cross-sectional study was conducted using a final-year MBBS exit examination comprising a Multiple-Choice Question (MCQ) test and an Objective Structured Clinical Examination (OSCE). Fifteen experienced faculty applied the Modified Angoff method to determine initial cut scores. Validity evidence was collected through pre- and post-judgment surveys, psychometric analyses, and acceptability surveys involving students, judges, and medical education experts. Descriptive statistics, standard error adjustments, and independent t-tests were used.
Results: Initial cut scores were 64.44/100 (MCQ) and 12.85/20 (OSCE). Judges showed high clarity, confidence, and agreement. Psychometric analysis revealed limited score variability, indicating consistent performance. Cut scores were adjusted upward by one standard error, a theoretically justified approach intended to reduce potential false-positive classifications. Acceptability was high among judges and students, with no significant group differences.
Conclusion: Post standard setting evaluation added validity evidence, reinforcing the credibility and acceptability of cut off scores in high stakes examinations.

Keywords: standard setting, modified Angoff method, cut-off score, high-stakes examination, Rwanda

Equlinet Misganaw Amare[1]
Abebe Bekele Zerihun[2]
Natali McCall[3]
Tefera Tadesse Jimma[1]
Hiba Abuelgasim Fadlelmoula Abdelrahman[1]
Anteneh Gadisa Belachew[2]
Biniam Ewneti Zelelew[2]
Olayinka Rasheed Ibrahim[3]

1 University of Global Health Equity, Division of Academic and Research Affairs, Education Development and Quality Center, Kigali, Rwanda

2 University of Global Health Equity, School of Medicine, Division of Clinical Medicine, Department of Surgery, Kigali, Rwanda

3 University of Global Health Equity, School of Medicine, Division of Clinical Medicine, Department of Pediatrics, Kigali, Rwanda

1. Introduction

Standard setting is the process of determining the level of knowledge and skills needed to achieve a typical performance level and then establishing a score on the exam scale that reflects that performance standard [1], [2], [3]. It is a systematic way of gathering value judgments, reaching consensus, and expressing that consensus as a single score on a test [3]. A cut-off score on a test is a special score that serves as the dividing line between those who perform adequately and those who do not [1], [3], [4].

These standards should ideally reflect the learning objectives and represent "mastery" of the material or a fundamental level of proficiency needed to advance to the next stage or function in real-world situations [5]. Standards for licensure and certification tests in medical education are designed to ensure that candidates have achieved the desired competencies for clinical practice [6], [7].

While different standard-setting methods often result in varying passing scores, due to the absence of a "gold standard", the key to establishing standards lies in selecting credible judges and employing systematic methods to gather their judgments [8]. The chosen approach should provide a cut-off score or validity evidence that enables precise categorization, reflects the examinee's performance, training, or instruction, and is statistically sound, practical, credible, acceptable, easy to implement, and straightforward to interpret [6], [7]. Standard setting is an essential part of assessment in health professions education, but determining a credible, defensible, and acceptable pass score remains a challenge, as no single method works perfectly [6]. Arbitrary pass marks are no longer advocated; therefore, systematic approaches are necessary to establish credible, acceptable, and defensible standards [8]. Various methods exist for establishing pass scores, including the Angoff, Ebel, Hofstee, Borderline group, and contrasting groups methods [8], [9], [10]. These methods rely on expert judgment and systematic approaches to capture their perspectives. The Angoff method, renowned for its inherent simplicity, is the most commonly used approach, with many institutions relying on it to establish pass scores for their students [10].

Regardless of the method used, a pass score should be supported by valid evidence, and it's essential to thoroughly evaluate the standard-setting process to ensure its validity, soundness, and defensibility [11]. Contemporary perspectives on assessment emphasize that validity is not an inherent property of a test but rather concerns the degree to which evidence and theory support the interpretation and use of test scores. Messick's unified validity framework conceptualizes validity as a comprehensive construct that integrates multiple sources of evidence, including content, internal structure, and consequences, within a coherent argument supporting score interpretation and use [12]. Similarly, Kane's argument-based approach to validation frames validity as the process of evaluating a series of inferences linking observed performance to decisions, supported by systematically collected evidence [12], [13].

Within these contemporary frameworks, validity evidence is typically derived from multiple sources, including:

1. content evidence, reflecting the representativeness of assessment tasks;
2. response process evidence, concerning the cognitive processes of judges or examinees;
3. internal structure evidence, such as variability and reliability indices;
4. relations with other variables; and
5. consequential evidence, including the impact of decisions and their fairness.

Importantly, these sources of evidence collectively contribute to a validity argument rather than functioning independently [14].

Despite this conceptual advancement, standard-setting practices in health professions education often focus primarily on the application of a recognized method, with limited attention to post hoc evaluation and to the systematic integration of multiple sources of validity evidence. There is increasing recognition that defensibility requires not only the application of established procedures but also the collection of supporting evidence on the process, outcomes, and consequences of cut-score decisions. In particular, examining expert judgment consistency, psychometric characteristics of cut scores, and stakeholder acceptability can provide important insights into the credibility of standard-setting outcomes [11], [15].

In this study, we adopted a systematic, multi-source approach to strengthen the validity and defensibility of cut-off scores derived using a modified Angoff procedure in a high-stakes medical examination. Drawing on contemporary validity frameworks, we explicitly examine multiple sources of evidence, including content (expert judgments), response processes (panel training and perceptions), internal structure (variability and standard error analyses), and consequences (acceptability among stakeholders). This study aims to evaluate the standard setting process by systematically generating validity evidence to support cut off scores. Through psychometric analysis, expert judgment, and post hoc review, the study seeks to enhance the credibility, defensibility, and acceptability of both the process and the resulting scores in high stakes medical examinations.

2. Materials and methods

2.1. Study setting and context

The study was conducted at the University of Global Health Equity (UGHE) in Rwanda. UGHE is a private, not-for-profit, accredited academic institution established in 2014 by Partners In Health (PIH), a non-governmental organization. It was founded in response to a critical global need – to train a new generation of health leaders capable of building and sustaining equitable health sys-

tems in underserved regions worldwide. UGHE offers a range of core programs, including a Master of Science in Global Health Delivery (MGHD), a dual program of Bachelor of Medicine and Bachelor of Surgery (MBBS) with MGHD, a Master of Science in Global Nursing Leadership, and a Master of Science in Health Professions Education. Each year, the MBBS program enrolls approximately 60 students through a competitive selection process, with around 70% of the cohort comprising female students. The program spans six and a half years, culminating in an exit examination at the end of the fifth year. The exit exam aims to determine whether students have met the minimum competency standards required to advance toward graduation and professional practice. The exam consists of a written cognitive test (200 multiple-choice questions) and a performance test using Objective Structured Clinical Examination (OSCE) stations (15 stations and three online ENTRUST examination platform), all developed according to a test blueprint. Once the exams were prepared, they were reviewed by a team of subject-matter experts and medical education specialists. A passing score for both the knowledge and performance tests was then established by a panel of experts (faculty members) using the Modified Angoff method before the tests were administered to the students. The target students for this study were the final-year MBBS students of the 2026 cohort.

2.2. Procedures used to evaluate the standard-setting process and generate valid evidence to support the resulting cut score. 15 stations plus three online ENTRUST examination platforms

2.2.1. First: Setting a cut-off score/pass score for the written cognitive knowledge – multiple choice questions (MCQs) and objective structured clinical examination (OSCE) performance test

To set the cut-off scores for cognitive knowledge and performance tests, we used a modified Angoff method. In this approach, the judges (expert faculty members) first defined the characteristics of a borderline group of examinees. They then assessed the difficulty and importance of each item or performance case on the test. Each judge estimated the percentage of hypothetical borderline examinees who would respond correctly to the item or performance case, and their estimates were recorded for group review. The judges discussed and revised their estimates as needed, including reaching consensus when the differences exceeded 0.2 points. The averages of the judges' estimates for each item or case were calculated, and the cut-off point was determined as the sum of those averages [1], [3].

2.2.2. Second: Collecting valid evidence to support the resulting cut-off score

While various approaches exist for gathering valid evidence to justify the pass score [4], this study employed the following methods based on feasibility.

2.2.2.1. Judges' survey

Surveys of the judges were conducted after the orientation and training session and after they completed the standard-setting exercise. The standard setting feedback questionnaire for judges [11] was used to evaluate whether the judges felt sufficiently trained for the process, their ability to make the necessary judgments, and their confidence in the resulting cutoff scores (see table 1).

2.2.2.2. Psychometric analysis

A structured approach was employed to evaluate the psychometric properties of the standard-setting outcomes and to systematically review the results of the standard-setting process [11]. The standard error of the mean (SE Mean) was calculated to estimate the precision of the derived cut score, and comparisons were made between the passing score and the distribution of examinees' test scores. In addition, the standard deviation (SD) of the judges' ratings was calculated to assess the degree of agreement among panel members, and this was also examined in relation to examinee score distributions [16]. The SE was computed as the standard error of the panel mean (SD/vn) and is explicitly distinguished from the standard error of measurement (SEM) associated with test reliability to avoid conceptual ambiguity. These indices were interpreted as providing supportive evidence of consistency and stability, rather than definitive proof of defensibility.

Following the initial estimation, the cut score was adjusted upward by one standard error. This adjustment was implemented as a normative, theoretically informed decision appropriate for high-stakes assessment contexts, where minimizing misclassification – particularly false-positive decisions (i.e., passing under-qualified candidates) – is a critical consideration [11], [17].

2.3. Surveying to evaluate how acceptable the cut-off score is

The acceptability of the standard-setting process and its resulting cut-off score is crucial and should be validated by all relevant stakeholders, including students, not just content experts or judges [11]. We gathered acceptability data from judges involved in the pass-score-setting process, medical education experts, and all students who took the exit exam from October to November 2025. For judges, acceptability data were collected both before and after the standard-setting process, allowing evaluation of changes in perceptions following participation in the procedure. For students, data collection occurred after

Table 1: Judges' ratings of the passing score-setting process after completing the standard-setting exercise

Item	Response category	n (%)
Difficulty of providing ratings	Not difficult	12 (80.0)
	Difficult	3 (20.0)
	Very difficult	0 (0.0)
Adequacy of time for the rating task	Too much time	1 (6.7)
	Right amount of time	14 (93.3)
	Not enough time	0 (0.0)
Adequacy of time for discussion	Adequate	12 (80.0)
	Inadequate	3 (20.0)
Appropriateness of cut scores	Too high	0 (0.0)
	Appropriate	13 (86.7)
	Too low	2 (13.3)
Confidence in cut scores	Very confident	14 (93.3)
	Confident	1 (6.7)
	Not confident	0 (0.0)

Table 2: Comparison of variability indices for cut scores and examinee scores

Test type	Format	Metric	Cut score	Examinee scores	Difference
Written test	MCQ	Mean	64.44	81.08	16.64
		SD	0.08	4.76	4.68
		SE	0.03	0.80	0.77
Performance test	OSCE	Mean	12.85	16.86	4.01
		SD	0.08	2.08	2.00
		SE	0.03	0.34	0.31

they were briefed on the process used to determine the cut-off score. The Acceptability of Intervention Measure (AIM), a five-point Likert scale (1="completely disagree" to 5="completely agree"), was utilized for this purpose. This tool is known for its strong psychometric properties and reliability in assessing acceptability [18] (see table 2). Data was collected from July to August 2025 using the Open Data Kit (ODK), an open-source Android app freely available for survey-based data collection purposes [19].

2.4. Data analysis

The ODK CSV data was imported into Statistical Package for the Social Sciences (SPSS) for cleaning and analysis. Descriptive statistics were used to describe participant characteristics and to summarize and visualize the data. Psychometric analyses, including mean, standard deviation (SD), and standard error (SE), were performed to compare the cut-off score with examinee scores. Additionally, to account for the consequences of different types of classification errors (false negatives or false positives), error minimization was conducted to address this concern. An independent t-test was performed to analyze differences in acceptability between two groups: students and expert judges (faculty members).

3. Results

3.1. Established cut-off scores for written knowledge and performance tests using the modified Angoff method

A total of 15 expert faculty members from four main departments – internal medicine, pediatrics, surgery, and obstetrics & gynecology – participated in setting the passing scores for both the written knowledge and performance tests. The process was facilitated by medical education specialists from the Education Development and Quality Center at UGHE. All participating judges were specialists and subject matter experts in their respective fields, having received orientation and training in the standard-setting process. Using the modified Angoff method, the judges established cut-off scores of 64.44 out of 100 for the written knowledge test and 12.85 out of 20 for the performance test. The standard deviations (SD) of panel ratings were relatively small across both components, indicating a high degree of consistency among panelists.

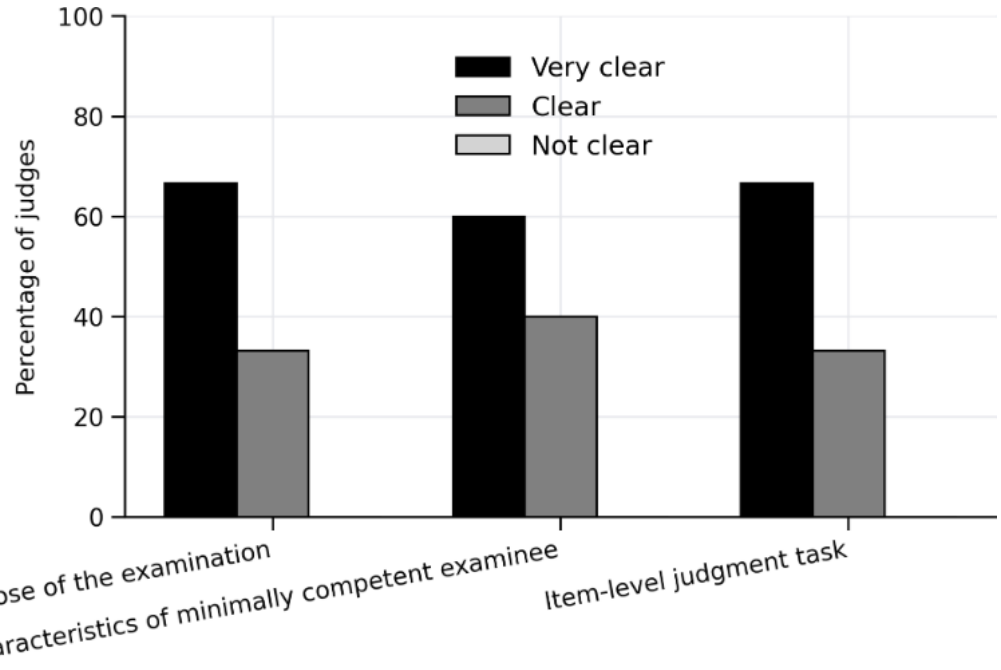


Figure 1: Judges' ratings of clarity regarding key components of the standard-setting process following orientation and training (n=15)

3.2. Judges' survey

A pre–standard–setting survey assessing the clarity of the process showed that all judges (100%) found it either very clear or clear with respect to the test's purpose, the characteristics of borderline examinees, and the rating tasks to be performed (see figure 1).

Similarly, after the standard-setting exercise, 80% of the judges reported that the process was not difficult. Regarding time allocation, the majority (93.33%) indicated that the rating task was allotted an appropriate amount of time, and all judges (100%) agreed that either the right amount or more than enough time was provided for discussion. Furthermore, all judges (100%) expressed confidence in the appropriateness of the final cut scores. Only 13.3% of the judges perceived the final passing scores as too low for examinees (see table 1).

3.3. Psychometric analysis

Psychometric analysis indicated that the standard deviation (SD) and standard error (SE) of the passing score for the written cognitive knowledge test (MCQ format) were 0.08 and 0.030, respectively. Similarly, for the performance test (OSCE format), the SD and SE were 0.077 and 0.029, respectively. In both examination components, the SE values were less than one-fourth of the SD, suggesting internal consistency in panel judgments (see table 2).

3.4. SE-based adjustment of cut scores

In response to concerns about the potential for false-positive classification in this high-stakes assessment context, the initial cut scores were adjusted upward by one standard error. Following this adjustment, the revised cut scores increased to 65.25 for MCQs and 13.19(out of 20) for OSCEs. (see table 3).

3.5. Impact of cut score adjustment

To evaluate the practical implications of the SE-based adjustment, pass rates were examined before and after the revised cut scores were applied. For the MCQ component, the pass rate at the initial cut score was 64.44%. Following application of the SE-adjusted cut score, the pass rate remained largely stable at 65.25%, indicating minimal change in classification outcomes at the group level. For the OSCE component, the cut score increased from 12.85 to 13.19 (out of 20) after the SE-based adjustment. This upward revision reflects a more stringent performance threshold; however, the magnitude of the adjustment was relatively small, suggesting limited impact on overall classification decisions

3.6. Student and judges acceptability survey

A total of 49 participants, comprising 15 judges, three medical education experts involved in the standard-setting process, and 31 students who took the exit examination, completed the pass score acceptability survey. The overall mean scores of judges, experts, and students reflect general agreement with the methodology used to establish the passing score and with the resulting cut score. (see table 4)

An independent two-sample t-test, assuming equal variances (F=0.64, p=0.28), revealed no statistically significant difference in the overall composite mean acceptability

Table 3: Original and adjusted cut scores

Test type	Format	Original cut score	SD	SE	Adjusted cut score (+1SE of examinee test)
Written test	MCQ	64.44	0.08	0.8	65.25
Performance test	OSCE	12.85	0.077	0.34	13.19

Table 4: Students' and judges/experts' ratings on the acceptability of the standard-setting process

Item	Students mean (SD)	Judges mean (SD)
Approval of process	4.71 (0.53)	4.33 (0.49)
Process is appealing	4.65 (0.49)	4.28 (0.67)
Liking the process	4.71 (0.46)	4.44 (0.51)
Welcoming the process	4.84 (0.37)	4.50 (0.71)
Overall mean	4.73 (0.46)	4.39 (0.59)

Table 5: Comparison of overall acceptability scores between groups

Group	n	Mean	SD	t (df=47)	p-value	95% CI
Judges/experts	18	4.39	0.59	0.60	0.55	[−0.21, 0.38]
Students	31	4.73	0.46			

scores between judges/experts and students (see table 5).

4. Discussion

This study is the first in Rwanda and the East, Central, and Southern Africa region to examine the defensibility of cut scores derived using a modified Angoff procedure by integrating multiple sources of evidence, including expert judgment, psychometric indicators, and stakeholder perceptions. The findings underscore the importance of adopting a multi-source approach to standard setting, particularly in high-stakes assessment contexts where decisions carry significant consequences [4].

The high level of confidence reported by panelists in the standard-setting process suggests that judges were adequately trained and capable of making informed decisions regarding cut scores [12]. It also indicates that faculty representation across disciplines was appropriate, ensuring that key perspectives on minimal competence were incorporated into the process. In this context, panelists' high confidence in the modified Angoff procedure provides content and response-process evidence, indicating that the procedure was perceived as clear, structured, and aligned with expectations for minimally competent performance [11], [20]. While this supports the credibility of the judgment process, such perceptions should be interpreted cautiously, as confidence alone does not guarantee the accuracy or validity of the resulting standards.

The relatively small standard deviation (SD) and standard error (SE) observed in this study provide evidence of internal structure, indicating convergence and consistency in panel judgments. These observations are broadly consistent with commonly cited heuristics in the standard-setting literature. For example, Jaeger (1991) suggests that the SE of the cut score should be less than one-fourth of the SD of examinee scores, while Meskauskas (1986) recommends that the variability of panel judgments (SD) should also remain below one-fourth of the examinee score SD [16], [21]. In the present study, these heuristic criteria were met for both the MCQ and OSCE components, suggesting internal consistency in panel ratings. However, these benchmarks are interpreted cautiously as rule-of-thumb indicators rather than definitive evidence of reliability or defensibility. Although the relatively small SD and SE values suggest convergence in panel judgments, they are descriptive indicators of agreement rather than direct measures of reliability and, in isolation, do not establish the robustness or validity of the derived cut scores. Accordingly, these psychometric indicators are interpreted as supportive but not conclusive evidence and are considered alongside other sources of validity evidence in evaluating the overall defensibility of the standard-setting outcomes.

The SE-based upward adjustment of the cut score represents a consequential consideration, reflecting a precautionary approach in high-stakes contexts where classification errors – particularly false positives – may have serious implications [11], [22]. Increasing the pass threshold aligns with recommendations to minimize the risk of certifying underqualified candidates [6], [23]. However, the modest changes observed in pass rates suggest limited impact on classification outcomes, indicating that the initial cut scores were broadly aligned with examinee performance.

Finally, the high level of acceptability among both judges and students provides additional consequential evidence, reflecting perceived fairness and credibility of the standard-setting process. While such perceptions enhance the legitimacy of the assessment, they should be interpreted

as indicators of stakeholder trust rather than objective validation of the cut scores. Taken together, these findings highlight that the defensibility of cut scores is strengthened through the integration of complementary sources of validity evidence, rather than reliance on any single indicator, and underscore the importance of cautious and balanced interpretation in high-stakes assessment contexts [20].

4.1. Strengths and limitations of the study

A key strength of this study lies in its pioneering role in systematically evaluating critical, yet often overlooked, components of the standard-setting process by applying rigorous methodologies and generating validity evidence that enhances the acceptability, defensibility, and credibility of both the process and its resulting cut-off score. However, a notable limitation is the lack of stakeholder validation. The study did not include perspectives from external groups such as policymakers or regulatory bodies, whose input is essential to affirm the broader acceptability and legitimacy of the standard-setting process. Future research should actively engage these stakeholders to build consensus and strengthen the generalizability of the findings.

5. Conclusion

The systematic approach used to generate validity evidence for the cut-off score contributed to its credibility, defensibility, and acceptability, as further affirmed by both students and faculty members. These findings underscore the importance of employing structured approaches in standard-setting to ensure fairness, trustworthiness, and stakeholder confidence in high-stakes assessments. Future studies should incorporate the perspectives of external stakeholders, such as policymakers and regulatory bodies, to further strengthen the legitimacy and broader applicability of the standard-setting process.

Abbreviations

- MBBS: Bachelor of Medicine and Bachelor of Surgery
- MCQ: Multiple choice questions
- ODK: Open Data Kit
- OSCE: Objectively structured clinical examination
- SD: Standard deviation
- SE: Standard error
- UGHE: University of Global Health Equity

Notes

Ethics approval and consent to participate

The study was approved by the Human Research Ethics Committee of the University of Global Health Equity Institutional Review Board. We adhered to all ethical guidelines for research involving human participants. Prior to data collection, an information sheet was used to obtain informed oral consent from participants. The consent form clearly explained the purpose, procedures, potential risks, and benefits of the study, ensuring that participants read and fully understood the information before agreeing to take part in the research.

Author contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Authors' ORCIDs

- Equilnet Misganaw Amare: [0000-0002-1115-6911]
- Abebe Bekele Zerihun: [0000-0002-0018-9096]
- Natali McCall: [0000-0001-9387-4521]
- Tefera Tadesse Jimma: [0000-0002-1770-4659]
- Hiba Abuelgasim Fadlelmoula Abdelrahman: [0000-0002-2568-9620]
- Anteneh Gadisa Belachew: [0000-0002-9836-2637]
- Biniam Ewnewti Zelelew: [0000-0002-3899-2880]
- Olayinka Rasheed Ibrahim: [0000-0002-2621-6593]

Acknowledgements

We are grateful to the University of Global Health Equity in Rwanda for the ethical clearance. We would also like to thank all the faculty and students who participated in this study.

Competing interests

The authors declare that they have no competing interests.

References

1. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach. 2008;30(9-10):836-845. DOI: 10.1080/01421590802402247
2. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 2014;36(2):97-110. DOI: 10.3109/0142159X.2013.853119
3. Norcini JJ. Setting standards on educational tests. Med Educ. 2003;37(5):464-469. DOI: 10.1046/j.1365-2923.2003.01495.x
4. Hejri SM, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. Med J Islam Repub Iran. 2014;28:34.
5. Hambleton RK, Van der Linden WJ. Advances in item response theory and applications: An introduction. Appl Psychol Meas. 1982;6(4):373-378. DOI: 10.1177/014662168200600401
6. Barman A. Standard setting in student assessment: is a defensible method yet to come? Ann Acad Med Singap. 2008;37(11):957-963.
7. Berk RA. A consumer's guide to setting performance standards on criterion-referenced tests. Rev Educ Res. 1986;56(1):137-172. DOI: 10.3102/00346543056001137
8. Yudkowsky R, Downing SM, Tekian A. Standard setting. In: Yudkowsky R, Park YS, Dowing SM, editors. Assessment in health professions education. London: Routledge; 2019. p.86-105. DOI: 10.4324/9781315166902-6
9. Hays R. Standard setting. Clin Teach. 2015;12(4):226-230. DOI: 10.1111/tct.12395
10. Swanwick T, Forrest K, O'Brien BC. Understanding medical education. Hoboken (NJ): Wiley-Blackwell; 2018. p.1-6. DOI: 10.1002/9781119373780
11. Yudkowsky R, Park YS, Downing SM. Assessment in health professions education. Vol. 26. London, New York: Routledge; 2019. DOI: 10.4324/9781315166902
12. Messick S, Linn R. Educational measurement. New York: American Council on Education; 1989.
13. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to K ane's framework. Med Educ. 2015;49(6):560-575. DOI: 10.1111/medu.12678
14. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):e166-e167. DOI: 10.1016/j.amjmed.2005.10.036
15. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 2000;22(2):120-130. DOI: 10.1080/01421590078526
16. Meskauskas JA. Setting standards for credentialing examinations: an update. Eval Health Prof. 1986;9(2):187-203. DOI: 10.1177/016327878600900205
17. Clauser BE, Mee J, Baldwin SG, Margolis MJ, Dillon GF. Judges' Use of Examinee Performance Data in an Angoff Standard□Setting Exercise for a Medical Licensing Examination: An Experimental Study. J Educ Meas. 2009;46(4):390-407. DOI: 10.1111/j.1745-3984.2009.00089.x
18. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, Boynton MH, Halko H. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci. 2017;12(1):108. DOI: 10.1186/s13012-017-0635-3
19. JeffreyCoker F, Basinger M, Modi V. Open data kit: Implications for the use of smartphone software technology for questionnaire studies in international development. 2010.
20. Norman GR, van der Vleuten CP, Newble DI. International handbook of research in medical education. Vol. 7. Berlin, Heidelberg: Springer Science & Business Media; 2012. (Springer International Handbooks of Education; 7).
21. Jaeger RM. Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' Assessments: A methodological accounting. J Pers Eval Educ. 1998;12(2):189-210. DOI: 10.1023/A:1008085128230
22. Clauser BE, Harik P, Margolis MJ. A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. J Educ Meas. 2006;43(3):173-191. DOI: 10.1111/j.1745-3984.2006.00012.x
23. Walter RA, Kapes JT. Development of a procedure for establishing occupational examination cut scores: A NOCTI example. J STEM Teach Educ. 2003;40(3):Article3.
24. Norcini JJ, Shea JA. The credibility and comparability of standards. Appl Meas Educ. 1997;10(1):39-59. DOI: 10.1207/s15324818ame1001_3

Corresponding author:
Equlinet Misganaw Amare
University of Global Health Equity, Division of Academic and Research Affairs, Education Development and Quality Center, Kigali, Rwanda
aequlinet@ughe.org

Please cite as
Amare EM, Zerihun AB, McCall N, Jimma TT, Abdelrahman HAF, Belachew AG, Zelelew BE, Ibrahim OR. Enhancing the defensibility of cut scores: A multi-source evaluation of a modified Angoff standard-setting process in a high-stakes medical examination. GMS J Med Educ. 2026;43(7):Doc95.
DOI: 10.3205/zma001889, URN: urn:nbn:de:0183-zma0018891

This article is freely available from
https://doi.org/10.3205/zma001889

Received: *2025-12-13*
Revised: *2026-05-28*
Accepted: *2026-08-10*
Published: *2026-11-16*

Copyright
©2026 Amare et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 License. See license information at http://creativecommons.org/licenses/by/4.0/.