Standard setting, a review of methods

Abstract:

In the field of education and assessment, cutoff scores are often used to distinguish between levels of achievement on a particular test or assessment. Standard setting is the process of establishing these cutoff scores, which are used to determine passing or failing grades, determine eligibility for certain programs, or identify individuals who may need additional support or intervention. Arbitrary cutoff scores are frequently utilized in many nations and institutions, and most commonly is the 60 % threshold. However, are students scoring around 60% considered competent and proficient? Is this passing threshold acceptable in professions where graduates will be treating and managing patients like Medicine and Dentistry? When implementing standard-setting techniques and when determining the right cutoff point, all of these issues need to be taken into consideration. Rather than using an arbitrary cutoff score of 60% as the passing level, it must be based on a methodical approach that adheres to a scientific technique with a history of validity and reliability. There are many methods which can be used to determine the cut off scores for different examinations. These can be based on judgements of borderline test takers, and these include for examples: Angoff, Nedelsky, and yes and no methods. In general, these methods depend on judges to determine characteristics of certain students who are defined as borderline, where judges evaluate test items before the examination, and assign numerical values for each item based on their belief whether these borderline students can answer these question items or not. Other methods may depend on judgements of examinees. These may include: borderline group, or contrasting group methods. These methods depend on the actual information about individual test takers including their test score and their level of knowledge and skills. Other methods to mention include: bookmark method, body of work, and compromise method.

Key words: Standard setting, Cut off score, Assessment, Angoff,

Introduction

In the field of education and assessment, cutoff scores are often used to distinguish between levels of proficiency or achievement on a particular test or assessment. Standard setting is the process of establishing these cutoff scores, which are used to determine passing or failing grades, determine eligibility for certain programs, or identify individuals who may need additional support or intervention [1].

The American Educational Research Association's Standards for Educational and Psychological Assessment [2], which represents the global agreement of the educational world on assessment, does not recommend the use of arbitrary cutoff values like the 60%. Yet, arbitrary cutoff scores are frequently utilized in many nations and institutions. Graduates in professions like medicine or dentistry will treat and oversee patients on daily basis; as a result, they might be held publicly accountable for their professional competency. Are students who score around 60% considered competent if they barely pass? How about the 40% of the test questions they were unable to respond to? Another thing to consider is how other factors, such as patient outcome, might fit into this.

When implementing standard-setting techniques and when determining the right cutoff point, all of these issues need to be taken into consideration. Rather than using an arbitrary cutoff score of 60% as the passing level, it must be based on a methodical approach that adheres to a scientific technique with a history of validity and reliability. Having a fixed norm would make it difficult to improve exams [3]. The issue may lie in the deliberate selection of content with a focus on a predetermined criterion, which undermines the reliability of judgments made as a result of these analyses.

Assessments are required of students in all academic disciplines to determine their proficiency in the subjects they have learned. This is especially relevant for occupations whereby trainees must be educated and accredited in their specific specialties. Since it is also directly related to patient safety, the necessity to certify trained persons lies at the core of many health professions [2]. To distinguish between candidates who are competent and those who are not, pass or fail cutoff points must be established [4]. As a result, calculating cutoff scores is crucial for high-stakes exams like licensure exams or tests required for further degrees, as well as for career and employment chances [5]. Consequently, a method and teaching tool are required to establish these thresholds. The process of gradually achieving justifiable cut scores through education is known as standard setting.

Review of literature

The two types of standard-setting methodologies are normative (relative) and criterion-based (absolute). Relative standards show how candidates stack up against a predetermined group; for instance, a group's score might fall short of the mean or higher than the standard deviation. The candidate's performance must meet a predetermined fixed passing score for the standards to be considered absolute [5].

Setting cut scores methods

I. Based on the judgments of borderline test takers

The idea is that the passing score of a test should be between the upper and lower test takers' scores, meaning that these takers' knowledge or performance is on the borderline between upper and lower takers. They were referred to as the "F-D" students by Nedelsky (1954)[6]. These approaches can be used prior or following the test. Administering these methods require obtaining information or data about the test questions' content by expert judges. Judges are often required to assess the probability of answering each test question by the defined borderline test takers. An auxiliary or supporting information about the test performance may be needed and can serve as reality check when setting the cut scores [7]. The steps required to perform these methods are similar and basically include: selecting the content experts (judges), defining borderline students, training the judges, collecting their judgments, then finally choosing the cut scores based on the judgments.

Nedelsky's method

Proposed by Nedelsky in 1954[6]. It may only be used for multiple choice questions. The premise of this method is that judges provide judgments about the probability of borderline test takers to correctly eliminate the wrong answers. This process may be slow and takes more time compared to other methods. Having all judges together is an advantage but it may end up taking a long time to reach consensus among raters or judges. If judges are going to rate the test individually, then it may be best to give them a sample of the questions to practice before discussing their answers with each other. Then they can proceed to rate the test individually. Collection of the data follows and the cut scores are subsequently calculated by averaging the judges' ratings.

Angoff's method

Proposed in 1971 by Angoff [8] and comparable to Nedelsky's method in that judges provide feedback on each test question. However, they do not provide the probability of eliminating wrong answers. Judges are obliged to determine the likelihood that the defined borderline examinees will correctly answer each test question. This method may thus be extended to be used with other test formats than multiple choice questions. Each judge assigns a score of 1 to each question if they believe the borderline examinee will correctly respond to it, and a score of 0 to each question they believe the borderline examinee will incorrectly respond to. The test cutoff score is then calculated using the sum of these numbers. This is the original method. A variation or modification of the method proposed by Angoff himself as a footnote is for each judge to give a probability score from 0.0 to 1.0 for each question. One way of making judgments about each question easier is to ask the judges to estimate how many borderline students, say 100, will correctly respond to this question. It is almost certain that Angoff's method is the most widely employed process for setting standards [9],[10]. It is also the most thoroughly researched method [11].

Yes/No method

Suggested by Impara and Plake (1997) [12] instead of assigning a probability number for each question between 0.1 and 1.0, judges are asked to select "yes" for each question they believe borderline test-takers would correctly answer and "no" for each question they believe borderline test-takers would incorrectly answer. Then, each Yes answer would account for a value of 1, while each No answer would have a value of 0, much like the original method. Numbers are then added and averaged among judges to get the cut score an in traditional Angoff. This method is easier than the probability one, takes less time, and easier for judges calibration and training.

Extension of the method

A variation of this method may be used for setting cut scores for essays. As suggested by Hambleton and Plake in 1995[13], judges provide a score estimate for borderline test takers for each essay question. For example, an essay question complete score is 10 marks; judges estimate that borderline test takers would probably score 4, (or 3 as determined by other judges). An average is then estimated. This approach could be used with conventional Angoff for multiple choice questions if the exam has both types of questions. After collecting the judgments, they are averaged and the cut score of the test is reached.

Ebel's method

This method differs from the previous two in that it is a two-staged process. In this method developed by Ebel in 1972[14], each judge classifies the questions into two groups based on their difficulty (hard, medium, and easy), and their relevance (essential, important, acceptable, and questionable). Then each judge assesses the probability of borderline students to answer each question in the different groups (for example, essential and easy group, questionable and hard). After the data collection, the average is estimated for each rater and then their scores are average to reach the cut score of the test. This method is time consuming and can be confusing to perform by the judges. Also, the cut score calculation is complicated by the grouping of questions. However, it is certainly an advantageous process because it includes judgments about the questions' relevance and difficulty levels, which is lacking in other methods of setting cut scores.

II. Based on Judgments about individual test takers

There are mainly two methods which are the borderline group and the contrasting groups. Unlike the judgments made based on a hypothetical borderline test taker group, these methods depend on the actual information about individual test takers including their test score and their level of knowledge and skills. Judges must be qualified and make judgments-of the knowledge and skills of test takers-which must reflect their true opinions about the test takers.

Borderline group method

The cut score is the mark that test takers with equivocal abilities and knowledge should aim for. Rather than pointing out which questions borderline students might or might not get right, this method depends on judges identifying borderline test takers [7]. Judges have a conversation to come up with a characterization of a test taker who is teetering the boundary between mastery and non-mastery. It is essential to determine borderline examinees after the description has been agreed upon. The median of that distribution (50th percentile), which would serve as the suggested cut score for the borderline group, would be determined by distributing the scores of the borderline examinees. The median is employed because excessively high or low numbers have a considerably smaller impact on it. Due to the likelihood that an examinee with a very high or very low score does not truly belong in the group, this aspect of the median is particularly crucial for the borderline group technique. [15,16]

Contrasting groups method

According to their assessments of their knowledge and abilities, the test-takers are split into two groups: qualified and unqualified. The cut score is established after the scores are separated into intervals and the proportion of test takers at each level is computed. By locating the intercept of normally distributed curves that represent the score distributions of the groups classified by their level of competence, the cut-off point is established. [7,17].

III. Other methods for setting cut scores

Bookmark method

The test questions are arranged from easiest to hardest in accordance with item response theory. Afterward, placing a "bookmark" at the level or place where test-takers on the edge can answer the hardest question correctly and the easiest question that they fail to answer [18]. The bookmark technique has the advantage of using all items' difficulty statistics, which renders it data-driven. Another benefit is the panelist confidence. [19-20]

Body of work method

As the name of this method suggest, judges in this method evaluate the test takers body of work. Multiple-choice questions are difficult to evaluate using this method, but essays are easy. A response booklet for each test taker is created and judges provide assessments and match the knowledge required at that level with the response booklet answers [21-22].

Compromise method

The best possible score as well as the lowest acceptable score and acceptable failure rate are decided by the judges. This approach must be used after the exam has been given because it depends on the test results. After displaying a chart with the aforementioned criteria, a cut score might be determined [23-25].

Conclusion

The process of setting standards involves defining the level of performance that is considered acceptable or desirable for a particular test, and then establishing the minimum score that a test-taker must achieve to meet that standard. Standard setting can help determine whether students have met certain learning objectives or whether they are ready to move on to the next level of instruction. It will ensure that assessment results can be used for their intended purposes and that tests are fair and unbiased. Without clear standards, it can be difficult to determine if a test measures the items it is designed to measure, or whether it is being influenced by factors such as test-taker background, cultural differences, or test-taking strategies.

Understanding different means of standard setting can help to ensure that assessments are valid and reliable. It is important to select the method that is most appropriate for the specific assessment in question. By using appropriate methods of standard setting, test developers can ensure that their assessments are measuring what they are meant to assess and that the scores obtained from those assessments are reliable indicators of performance thus increasing transparency and fairness in assessment. Despite its importance, standard setting has also been the subject of controversy and debate in education. Critics argue that standardized tests and the standards they are based on can be biased, culturally insensitive, or not aligned with the needs and goals of students and communities. As a result, there has been a growing movement towards alternative forms of assessment and evaluation that take a more holistic and student-centered approach to learning and evaluation.

Competing Interest:

The author declares that there are no financial or personal relationships with other people or organizations that could inappropriately influence this work.

References:

1 Cusimano, M. D. Standard setting in medical education. *Academic Medicine*, 1996;*71*(10), S112-20.

2- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). 2014; Standards for educational and psychological testing. Washington, DC: AERA.

3- Tekian A, Norcini J. Overcome the 60% passing score and improve the quality of assessment. GMS Zeitschrift fur Medizinische Ausbildung 2015;32(4):1-2

4-Downing S, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examination in health professions education. Teaching and learning in Medicine 2006;18(1):50-57

5-Ben-David MF. AMEE guide No. 18: Standard setting in student assessment. Med Teacher 2000;22(2):120-130

6-Nedelsky L. Absolute grading standards for objective tests. Educational and Psychological Measurement 1954;14:3-19

7- Zieky, M, Perie M. A Primer on setting cut scores on tests of educational achievement. 2006;NJ: Educational Testing Service, Inc.

8- Angoff, WH. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement 1971; (2nd ed., pp. 508-600). Washington: American Council on Education

9- Cizek GJ, Bunch MB. The Angoff method and Angoff variations. In: Standard setting. A guide to establishing and evaluating performance standards on tests 2011; (pp. 81-95). Thousand Oaks, CA: Sage Publications

10- Meara KP, Hambleton RK, Sireci SG. Setting and validating standards on professional licensure and certification exams: A survey of current practices. Clear Exam Review 2001;12(2):17-23

11- Mills CN, Melican GJ. Estimating and adjusting cutoff scores: Features of selected methods Applied Measurement in Education 1988;1:261-275

12-Impara JC, Plake BS. Standard setting: An alternative approach. Journal of Educational Measurement 1997;34:353–366

13- Hambleton RK, Plake BS. Using an extended Angoff procedure to set standards on complex performance assessments Applied Measurement in Education 1995;8:41-56

14 Ebel, RL. Essentials of educational measurement (2nd Ed.)1972; Englewood Cliffs, NJ: Prentice-Hall.

15-Jaeger, R. M. Certification of student competence. In R. L. Linn (Ed.), Educational measurement 1989; (pp. 485–514). Macmillan Publishing Co, Inc; American Council on Education.

16-Livingston, S. A., & Zieky, M. J. A comparative study of standard-setting methods. Applied Measurement in Education, 1989;2(2), 121-141.

17-de Montbrun S, Satterthwaite L, Grantcharov TP. Setting pass scores for assessment of technical performance by surgical trainees. Br J Surg. 2015;103:300–306.

18- Mitzel HC, Lewis DM, Patz RJ, Green DR.The bookmark procedure: Psychological perspectives. In Edited by G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives 2001; (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

19-Lewis, D. M., Mitzel, H. C., & Green, D. R. Standard setting: A Book-mark approach. In D. R. Green (Chair),IRT-based standard setting procedures utilizing behavioral anchoring 1996. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

20-Karantonis, A., & Sireci, S. The Bookmark Standard Setting Method: A Literature Review. Educational Measurement Issues and Practice 2006;25(1):4 – 12.

21- Kingston NM, Kahl SR, Sweeney K, Bay L. Setting performance standards using the body of work method. In Edited by G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives 2001; (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum.

22-Wyse, Adam E.; Bunch, Michael B.; Deville, Craig; Viger, Steven G. A Body of Work Standard-Setting Method with Construct Maps. Educational and Psychological Measurement,
2014;74 (2):236-262

23- Hofstee, WK. The case for compromise in educational selection and grading. In Edited by S. B. Anderson & J. S. Helmick (Eds.), On educational testing 1983 ;(pp. 109–127). San Francisco Jossey-Bass.

24-Burr, S.A., Whittle, J., Fairclough, L.C. et al. Modifying Hofstee standard setting for assessments that vary in difficulty, and to determine boundaries for different levels of achievement. BMC Med Educ 2016;16, 34

25-Beuk, CH. A method for reaching a compromise between absolute and relative standards in examinations. J Education Measurement 1984;21(2):147-152

Appendix 1: Summary of different types of Educational and Psychological Assessment standard setting approaches

Approach	Advantages	Disadvantages	Uses
Nedelsky Method	Can help identify appropriate cut scores for a range of different assessments, including those with multiple domains.	Requires significant technical expertise to set up and interpret; may not be appropriate for all types of assessments.	Used in educational and professional contexts where there are multiple domains to assess, and the goal is to set appropriate cut scores for each domain.
Ebel's method	Facilitates transparency and fairness. Considers both the difficulty level of the items and the content coverage.	Can be difficult to use for assessments with large item pools. May require a high level of technical expertise to implement and interpret the results.	Commonly used approach for setting standards in educational and certification contexts
Angoff Method	Relatively easy to use; encourages consensus among subject matter experts.	Time-consuming to set up; can be difficult to identify the correct passing score.	Commonly used in educational contexts, such as licensure exams or other standardized tests.
Yes/No Method	Requires minimal resources and technical expertise. Encourages discussion and consensus among the panelists, leading to greater transparency and acceptance.	Can be prone to bias or errors in judgment, as it relies on the subjective opinions of the panelists. May not be suitable for all types of assessments, particularly those with a large number of items or complex domains.	commonly used in educational settings to set standards for high-stakes assessments, such as licensure or certification exams. It is also used in professional contexts to establish minimum competency levels for various professions.
Bookmark Method	More efficient than Angoff method; less dependent on expert judgment.	May not accurately reflect true ability levels; assumes that all test questions are of equal difficulty.	Used in educational settings, especially for large-scale testing programs like state-mandated assessments.
Borderline Regression	Statistical approach that considers multiple test items and their relative difficulty levels.	Requires a large sample size and significant technical expertise to set up; may not be appropriate for certain types of assessments.	Used in educational and professional settings where there are multiple test items of varying difficulty, and the goal is to set a pass/fail score.