Debate Article
http://mjiri.iums.ac.ir

Medical Journal of the Islamic Republic of Iran (MJIRI)
Iran University of Medical Sciences

Standard setting in medical education: fundamental concepts and emerging challenges

Sara Mortaz Hejri[1], Mohammad Jalili[2]

Received: 5 Jan 2014 Accepted: 3 Mar 2014 Published: 19 May 2014

Abstract

The process of determining the minimum pass level to separate the competent students from those who do not perform well enough is called standard setting. A large number of methods are widely used to set cut-scores for both written and clinical examinations. There are some challenging issues pertaining to any standard setting procedure. Ignoring these concerns would result in a large dispute regarding the credibility and defensibility of the method. The goal of this review is to provide a basic understanding of the key concepts and challenges in standard setting and to suggest some recommendations to overcome the challenging issues for educators and policymakers who are dealing with decision-making in this field.

Keywords: Student assessment, Standard setting, Reliability, Validity.

Cite this article as: Mortaz Hejri S, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. *Med J Islam Repub Iran* 2014 (19 May). Vol. 28:34.

Introduction

Student assessment is an integral part of educational programs. Since it drives students' learning and highlights significant goals and objectives of the course, teachers and administrative pay careful attention to its different parts. However, standard setting is an area in the field of assessment which is not dealt with so frequently.

A standard, also known as the minimum pass level, separates the competent students from those who are not. The process of determining this special score is called standard setting (1). The decision to pass or fail an examinee is an important issue in medical education, especially for licensure and credentialing purposes (2). The standard should not be set in an arbitrary way but it should be established through a specific methodology that considers the test's objectives and content areas, the examinees' performance, and the wider social or educational setting (3).

A large number of methods have been developed and used to set standard for both written and clinical examinations (4). Standard setting methods, depending on the purpose of the test, can be either norm-referenced or criterion-referenced. Norm-referenced (relative) standard setting methods are used when a fixed proportion of examinees are required to pass. Since the standard is based on the ability of the cohort of students, it is possible that some competent candidates would fail the exam. The criterion-referenced (absolute) methods, such as Angoff or borderline regression, deal with the desirable competency level that each student should achieve. So, hypothetically, all examinees may pass or fail a test with an absolute standard (2).

Each of the methods serves a particular

1. MD, MSc, PhD student, Medical Education Department, Tehran University of Medical Sciences, Tehran, Iran. sa_mortazhejri@razi.tums.ac.ir
2. (**Corresponding author**) MD, Associate Professor, Emergency Medicine Department, Medical Education Department, Tehran University of Medical Sciences, Tehran, Iran. mjalili@tums.ac.ir

purpose and none is agreed upon as the best method or gold standard for all settings (5). Many published studies have addressed this topic by delineating practical steps of various standard setting procedures. Furthermore, literature abounds with papers reporting the application of different standard setting methods and comparing their results in terms of obtained cut score, pass rates and the degree of error in the process. Providing a detailed description of the existing techniques is beyond the scope of this manuscript and can be found elsewhere in the medical education literature (6-11).

The goal of this review is to provide a better understanding of standard setting for educators and policymakers who are dealing with decision making in this field by focusing specifically on the challenging issues surrounding this topic. We will also discuss some possible solutions and suggestions to overcome these problems, hence obtaining more credible results.

Areas of concern

While each of the standard setting procedures possesses their unique specifications, they all share some challenging issues which might occur to anyone who is engaged in standard setting. Ignoring these concerns during the procedure may result in a large dispute regarding the credibility and defensibility of the method (3,5,12). These challenging issues include, but are not limited to, the following: the subjective nature of the standard setting, the definition of a minimally competent student, and the variability in standard setting results.

The subjective nature of the standard setting

One of the very first challenges in setting standards is that all of the methods require the application of "judgment" (13,14). In some methods, experts are asked to estimate the probability that a borderline candidate would correctly answer test items. Others require judges to observe and evaluate students' performance during the examination. In both procedures, the central and important role of judgment cannot be ignored (4). Because standards are an expression of subjective values, critics claim that they are not valid. It is important to consider, however, that no purely objective method for determining the cut-score exists (13). In other words, although particular statistical and mathematical methods are used as part of some standard setting approaches, there are no true cut-scores that can be achieved through application of a perfectly objective method. It should also be noted that human judgment plays a fundamental role in every level of student assessment and not merely in standard setting (14). Some of the issues reflecting the judgments of test takers include choosing type of item, establishing what questions to ask, writing and editing questions, selecting the best option in cued questions, and scoring constructed-response questions. It seems that the role of judgment in test development is accepted without difficulty while concerns about the subjective nature of standard setting are overemphasized.

The definition of a borderline student

Another important challenge in standards setting is the definition of the "borderline" student. Although application of this concept is more pronounced in some methods such as Angoff, in which judges should envisage a borderline candidate and estimate their performance, understanding the characteristics of such a student, is the cornerstone of almost all methods. It is frequently stated that the cognitive task of considering a borderline candidate is highly demanding even for the experts, to a degree which may impair their judgments. This is especially true if judges' concepts change from one item to another according to discussions or mental fatigue throughout the process (3). It has been noted that judges, in an effort to facilitate the creation of this conceptual image, think about an average student instead of focusing on the borderline performer, leading to the substitution of a criterion-

based concept with a norm-referenced one (13).

This issue is closely related to the general decision on students' proficiency levels. The classification of students' performance may be limited just to competent or incompetent, or might be labeled into 5 or 6 categories, each designating a certain level of competency with borderline performance lying somewhere in the continuum (13-15). While there is no universally-agreed rule for the number and definition of these levels, serious problems arise when judges try to explain the borderline category and justify its location on the scale.

Variability of cut-scores

Another criticism aimed at the credibility of standard setting is the variability of obtained standards. As the literature reveals, variability in standard-setting results using different techniques, or even across replications of the same procedure, can be large (7-11), adding weight to the argument that these methods cannot be trusted to distinguish competent students from non-competent candidates. Generally, when pass/fail decisions are made in an examination, two kinds of errors may lead in misclassification of students: the error associated with the test score and the error related to the determined standard (3). In fact, variability in observed scores can occur in any kind of repeated measurement and it is not limited to standard setting (14). It is not unusual for a student to take two so-called parallel exams and achieve two different scores. Nichols et al. argue that although both standard setting and student assessment lay in the field of measurement, they are not exactly the same. While the former should be regarded as a stimulus -centered approach, in which higher reliability will be obtained if the variance associated with items is large and the variance associated with persons is small, the latter is often treated like a subject-centered approach in which the higher reliability will be obtained if the variance associated with persons is large and the variance associated with items is small (14).

Suggestions for improvement

While the above-mentioned challenges are inherent to the procedure, several suggestions may reduce the concerns and enhance the outcome. Some of these recommendations need to be followed before setting the standard and some should be applied afterwards. Most of them can be adapted irrespective of the method selected for determining the cut score.

Selection of appropriate judges

The number and nature of the judges are central to the credibility of the standard. Judges have different cut scores in mind due to difference in their educational background, professional role, socioeconomic status, as well as their knowledge, experience, and opinions relating to the standard setting method (5,12,13).

In Angoff, Ebel, and Nedelsky, where formation of a panel of specialists is required, involvement of an appropriate number and mixture of the judges to include a variety of viewpoints and to generate acceptable results, is of paramount importance (12,13).

Although the exact number of the panelists required is still controversial and studies have yielded results as low as 5 and as high as 20, most suggestions revolve around a group of 10 judge as suitable for this purpose (16-18). Furthermore, factors such as the method of standard setting, the content area of the exam, and the presence (or absence) of group discussion or reality checks vary among these studies, limiting the generalizability of their findings.

The judges should also be good representatives of the relevant experts and should be selected meticulously, considering their age, gender, ethnicity, and educational experience.

Defining performance level and characteristics of a borderline student

Before a method is selected, the stake-

holders, that may or may not be different from the judges, should decide on students' performance levels including number of categories, their labels, and a behavioral descriptor for each category (13-15). Since most methods require judging the performance of a borderline student, development of criteria relating to minimally accepted competency is an important step. Detailed descriptors should demonstrate the knowledge, skills, and abilities in a specific context that are expected from a candidate in that category.

Training of the judges

Training the judges on the selected method, including the opportunity for practice, discussion, and feedback, is critically important. Bearing in mind the second challenge, it is essential to provide judges with the performance levels descriptors, and then let them reach a deep understanding through discussion with other panelists (3,14). Characterizing the borderline students by creating a list of relevant skills measured in the test, can help judges to reach a consensus (19).

Assessing the reliability of standard setting

As mentioned earlier, variability in cut scores obtained by different standard setting methods or on different occasions is inevitable. A frequently used framework to interpret this variability is the reliability or consistency of the results. As reliability estimates are used to acknowledge and delineate the magnitude of the error inherent in student assessment, a similar approach can be adapted to quantify the error component of the cut-score. In other words, by replication of the procedure or conducting another method or using another panel of judges, how consistent the cut-score would be or what proportion of students would be classified similarly. The more reliable a method, the less likely the results will be affected by large random errors.

Reliability can be calculated using Classical test theory (CTT) or Generalizability theory (GT). Under CTT, an observed score on a measurement is the sum of the true score and the error component. Sources of error in standard setting include different panelists, different context, and different occasions in which judgments occur (14,20). In contrast to CTT, which considers error to be unitary, GT can determine the contribution of all sources of variance at the same time. The intent of a G-study in this context is to differentiate among items while generalizing results over judges. But caution must be exercised in interpreting the reliability coefficient since it might be influenced by one judge who dominates others or endorses a shared misconception among panelists (5,6). In this way, higher reliability coefficients no longer reflects judges' true perceptions or expectations.

It should, however, be noted that reliability does not tell us about the meaningfulness of the standard and does not guarantee its appropriateness for the given purpose. This issue will be dealt with in greater detail in the forthcoming paragraphs.

Ensuring the validity of standard setting

The standard setting aims at dividing candidates into mastery and non-mastery categories and the validity of standard setting, also known as the credibility, deals with how well this task has been accomplished. A procedure that misclassifies a non-competent student as competent (false positive) or vice versa lacks accuracy.

One method to measure the validity of standard setting is to follow the students' performance in future. If the competent students show acceptable behavior in their workplaces, the standard will prove credible. However, in this design, it is impossible to compare the performance of competent and non-competent students because the latter are usually not permitted to pursue practice. Another method is to compare pass/fail rates of one test with that of other concurrent exams.

It is important to keep in mind that the

above-mentioned approaches do not prove the validity of the standard itself (i.e. 45 or 55 or …) since the 'true' cut-score does not exist. We might at best try to ensure that the chosen method is appropriate and can give rise to sound decisions. The appropriateness of the method is also supported when evidence of defensible process is demonstrated. It is evident that setting standards by gathering the judgments of a group of experts in an unbiased way, and with consideration of the level of the examinees and the content of the exam, makes more sense than relying only on a fixed pre-defined arbitrary score. For this reason, careful documentation of the whole process, including number and characteristics of experts, as well as collecting comments of judges and stakeholders about credibility of the results, should be considered. However, it should be noted that an appropriately set standard may make the pass/fail decisions defensible, but there is no conclusive way to ensure the validity of any standard-setting method and relying only on procedural evidence, provides weak justification for the credibility of the decisions.

Conclusion

Standard setting in the medical profession is still in an evolutionary stage. While various approaches have been developed, there are still many concerns regarding this process. Although these challenges cannot be fully eliminated, ensuring the quality of the standard setting, which can be accomplished by taking some of the steps mentioned in this manuscript, is of paramount importance. The information obtained through this quality assurance may be helpful in interpreting the standards and can also prove that, in spite of the variability of scores, the pass/fail decisions are defensible and reasonable.

Acknowledgement

We want to express our special gratitude to Professor John Norcini for his invaluable comments on the final draft of the manuscript.

References

1. Cusimano MD. Standard setting in medical education. Acad Med 1996; 71:112–20.
2. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher 2000; 22(2): 120-130.
3. Ricker KL. Setting Cut-Scores: A Critical Review of the Angoff and Modified Angoff Methods. The Alberta Journal of Educational Research 2006; 52(1): 53-64.
4. Cizek GJ, Bunch MB. Standard setting: a guide to establishing and evaluating performance standards for tests. Thousand Oaks, CA: Sage Publications, Inc., 2007.
5. Barman A. Standard setting in student assessment: is a defensible method yet to come? Ann Acad Med Singapore 2008; 37(11): 957-63.
6. Hurtz GM, Auerbach MA. A meta-analysis of the effects of modifications to the Angoff method on cut-off scores and judgment consensus. Educ Psychol Meas 2003; 63:584–601.
7. Wood TJ, Humphrey-murto SM, Norman GR. Standard Setting in a Small Scale OSCE: A Comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Adv Health Sci Educ Theory Pract 2006; 11(2): 115-22.
8. Jalili M, Hejri SM, Norcini J. Comparison of two methods of standard setting: the performance of the three-level Angoff method. Medical Education 45 (12), 1199-1208.
9. Jalili M, Mortaz Hejri S. Standard Setting for Objective Structured Clinical Exam Using Four Methods: Pre-fixed score, Angoff, Borderline Regression and Cohen's. Strides in Development of Medical Education 2012; 9 (1).
10. Mortaz Hejri S, Jalili M, Labaf A. Setting Standard Threshold Scores for an Objective Structured Clinical Examination using Angoff Method and Assessing the Impact of Reality Checking and Discussion on Actual Scores. Iranian Journal of Medical Education 2012, 885-894.
11. Mortaz Hejri S., Jalili M, Muijtjens AMM, Van der Vleuten CPM. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci 2013;18:887-91.
12. Norcini JJ. Setting standards on educational tests. Med Educ 2003; 37(5): 464-9.
13. Zieky M, Perie M. A primer on setting cut scores on tests of educational achievement. Princeton, NJ: Educational Testing Service, Inc. 2006
14. Nichols P, Twing J, Mueller CD, O'Malley K. Standard-Setting Methods as Measurement Processes. Educational Measurement: Issues and Practice 2010;29(1):14-24.
15. Chinn RN. Considerations in setting cut scores. Lexington, Kentucky: Council on Licensure, Enforcement, and Regulation, Resource Brief. 2006.
16. Brennan RL, Lockwood RE. A comparison of

the Nedelsky and Angoff cutting score procedures using generalisability theory. Appl Psychol Measurement 1980;4:219–40.

17. Hurtz GM, Hertz NR. How many raters should be used for establishing cutoff scores with the angoff method? A generalizability theory study. Educational and Psychological Measurement 1999; 59(6):885-897.

18. Fowell SL, Fewtrell R, McLaughlin PJ. Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? Adv Health Sci Educ Theory Pract 2008;13(1):11-24.

19. Carlson J, Tomkowiak J, Stilp C. Using the Angoff Method to Set Defensible Cutoff Scores for Standardized Patient Performance Evaluations in PA Education. J Physician Assist Educ 2009;20(1):15-23.

20. Alvaro J. Arce, Ze Wang. Applying Rasch Model and Generalizability Theory to Study Modified Angoff Cut Scores. International Journal of Testing 2012;12(1):44-60.