Standard setting in written assessment

Wadi, Majed M.

Published in:
Written Assessment in Medical Education

DOI:
10.1007/978-3-031-11752-7_10

Publication date:
2023

Licence:
Other

Document Version
Peer reviewed version

Chapter Ten: Standard Setting in Written Assessment

Majed Wadi

Medical Educationist and coordinator of the Assessment Unit, College of Medicine, Qassim University, Saudi Arabia.

About this chapter

The term "standard setting" refers to the process of establishing the minimum passing score for students on a test. It is not simply a matter of arbitrarily establishing a cut score for a test; rather, it is a laborious process by which a panel determines the cut score for a specific test in a particular context.

In medication education, certifying medical students as a doctor is very critical. The decision made to graduate health care practitioners should be based on rigorous methods to determine how much graduated doctors are safe for people's lives. For this reason, standard setting lay at the heart of the assessment. This chapter illuminates the standard-setting process and discusses pertinent methods used in medical education, particularly for written tests.

By the end of this chapter, the reader is expected to be able to

1. Recognize concepts in standard setting and their importance in Medical and Health Professions Education
2. Explain the common methods of standard setting in written assessment.

Overview of Standard Setting

The noble goal of medical education is to graduate competent and safe physicians. As a consequence, the results of certifying safe physicians should include rigorous validity evidence on the methods used to determine graduate competencies. The standard setting process lay at the heart of these evidences (1).

The standard setting means the process of creating a cut point or boundary to ascertain examinees into either passed, failed or borderline. To create this boundary, many efforts should be taken to. It is a crucial step in student assessment as a decision based on the standard has the potential effect not only on the careers

of examinees but also, and more importantly, on the lives of those who would benefit from examinees certified as competent (2).

Approaches for assigning a pass/fail status in an evaluative setting

There are two methods for classifying examinees as pass or fail: norm and criterion references. The following subheading highlights these techniques.

Criterion-referenced approach

A criterion-referenced standard is an absolute standard that is calibrated against a particular level of examinee performance or against a standard, predetermined competencies on a particular examination. Each examinee is evaluated in relation to this absolute standard, regardless of the examinee group's performance on that examination (3). In a nutshell, the methods discussed in this chapter are a form of criterion-based standard setting.

Norm-referenced approach

The performance of an examinee is evaluated in comparison to the performance of the entire group, rather than on its own merits, which is why it is referred to as norm reference. A norm-referenced standard is one that is established relative to the performance of a group of examinees on the same examination. Thus, the standard varies according to the examinee group's performance. This may result in misinterpretation in some instances. For example, an examinee placed in a group of examinees with a low performance standard has a better chance of meeting the standard than an examinee placed in a group with a high-performance standard.

While the process of developing a norm-referenced standard is much simpler than developing a criterion-referenced standard, there is no guarantee that the

standard will be equivalent between examinations, as examinee group performance may differ between examinations and cut-off scores are determined by group score distributions. Cut-off points based on norm-referenced standards are irrelevant when determining an examinee's competence or incompetence (3).

Common concepts in Standard Setting

Standard setting:

It is the methodology that is run by a panel to determine the minimum pass level (cut score) for a given test (3, 4).

Minimum pass level (MPL):

It is numerical output of the standard setting (number cut point) or boundary to ascertain examinees into either passed, failed or borderline (minimally competent student) (5).

Minimally competent (borderline) student:

A minimally competent or borderline student is one who is just on the border of failing. This student’s knowledge-base borders on the edge between competence and incompetence. the criteria for classifying students as borderline depend on several factors in a given context. These factors should be specified by the panel, for whom standard setting is assigned.

Panel:

It is a group of medical teachers acting as judges and area experts to determine the borderline students and set the minimum pass level accordingly (4).

General classification of standard setting methods:

- *Test-centered standards* are those derived from hypothetical decisions based on the test content. In these methods a group of expert judges set the standard by reviewing the items in the test and deciding on the level of examinee performance on these items that will be considered just adequate for demonstrating competence. Methods included in this category are:
  - the Angoff method (6),
  - the Nedelsky method (7), and
  - the Ebel method (8).

- *Examinee-centered standards* are those derived from reviewing the performance of examinees or a similar group prior to making judgments about what constitutes borderline performance between competence and incompetence. Methods included in this category are:
  - the borderline group method (9) and
  - the contrasting group method (9).

Common Methods of Standard Setting in written assessment

The Angoff (1971) method

The Angoff method is commonly regarded as the most popular method of standardization. It's appropriate for both written and performance assessment. This method is based on defining and determining the features of a "borderline" examinee, i.e. a marginally competent person on the edge of passing or failing. This activity should be done by a group of experts or seniors who are familiar with the specific population of students for whom the standard setting has been established. There are many variants of this method. However, for the purpose of this book and

to make this method simple and understandable, we will discuss the classical Angoff method.

Steps of Angoff Method:

1. Defining the borderline students.

A panel of judges will first gather to discuss the qualities of a "borderline" examinee, that is, someone who is marginally competent but on the edge of passing or failing. They should be told to think of an example from the intended students' population.

2. Review and rate test items

The first item is read aloud by one of the judges to begin the item review. The reader, followed by the other panel members, estimates how well a borderline applicant will perform on that item. Each judge is asked to consider a sample of minimally competent persons, say 100, and estimate the proportion of these individuals who would properly answer the item. Note that the difference between "will perform" and "should perform" needs to be stressed and considered by the panel. Each new item is judged in a clockwise rotation.

3. Record the rate:

For each item, it is possible to record the rating either by hand or with the use of an excel sheet or even Google's online form.

4. Review the rating and coming to consensus:

If considerable gaps (more than 20%) exist between the judges' judgments after they have all given their separate judgments on all of the test's questions, a group discussion may be performed to attempt to explain why such large variances exist. Judges now have the option to amend their earlier decisions independently if they so desire.

5. Calculate the cut-of score point

Using the Excel sheet is the appropriate and easy way to calculate the cut-off point. It is done by calculating the average of means of all test item across all raters (judges). Table () is an example. This average represents the cut-off point for making pass/ fail decisions.

Table 10-1. Record of test items rating by the panel

Questions	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Mean
Item 1	50	60	55	65	60	48.5
Item 2	75	80	70	85	77	64.8
Item 3	90	85	85	95	80	73.0
Item 4	55	50	55	60	45	44.8
Item 5	70	80	77	70	65	61.2
Average of means			58.5*			

*This the cut-off point (minimum pass level)

Nedelsky (1954) method

This method was developed for multiple-choice questions, in which a panel evaluates each MCQ item and its distractors and assigns a score to each option based on how frequently minimally competent (borderline) students choose each option. The average rating for each MCQ item is calculated, and then the average rating for all items is added.

Steps of Nedelsky Method:

1. Create a rating form (you can use an online form) that includes the serial number of the MCQ and its associated option, as well as a method for marking the key answer (using star for example).
2. Create a new file that contains the MCQs items (either word or PPt file)
3. Assemble a panel (5–10 senior medical educators who are subject-matter experts).
4. Brief the panel on the specific task assigned to them.
5. Define the minimally competent (borderline) student's criteria.
6. Display each MCQ item and request that the panel open the form.
7. Begin the session by asking the panel to rate each option as 1 if it is difficult to get the borderline student's attention and 0 if it is easy to get his/her attention.
8. The critical response should be designated as 1.
9. Add the sum of the distractors assigned as 1 to each MCQ item.

10. Estimate the MPL for each item using the following formula: MPL is equal to 1/m (m is the number of the distractors that were assigned as difficult)
11. Finally, a grand average for all items is calculated by adding the MPL of each item and dividing it by the total number of items.

12. **Table 10-2.** Example of Nedelsky method

Items	Options	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Average
Item 1	Option a	0	1	1	1	0	0.38
	Option b*	1	1	1	1	1	
	Option c	1	0	1	0	1	
	Option d	0	0	1	1	1	
	1/difficult options	1/2	1/2	1/4	1/3	1/3	
	Minimum pass index	0.5	0.5	0.25	0.33	0.33	
Item 2	Option a	1	1	1	0	1	0.36
	Option b	1	0	0	1	1	
	Option c*	1	1	1	1	1	
	Option d	0	1	0	1	0	
	1/difficult options	1/3	1/3	1/2	1/3	1/3	
	Minimum pass index	0.33	0.33	0.50	0.33	0.33	

13. *key answer

Practical Application

If we have a test with 30 MCQs, the Minimum pass index of each item should be added.

Example: Minimum pass index of item 1 + Minimum pass index of item 2 + … + weight of item 30

Suppose that we got the value of 22.98 for all 30 items.

The cut-off point (the minimum pass level MPL) for the Nedlesky method is then calculated by dividing MPI by the number of MCQs.

The MPL = 22.98/30 * 100 = 76.6%

The Ebel (1972) method

Two characteristics of each item should be considered in the Ebel method (Ebel 1972): difficulty (easy, medium, difficult) and relevance (essential to know, important to know, acceptable or nice to know). The judges then estimate the proportion of borderline examinees who will be able to respond to these types of questions (after their classification in two dimensions matrix).

Steps of Ebel Method:

Two tasks in the Ebel method should be completed sequentially. The first task is to classify test items according to two dimensions: difficulty and relevance; the second task is to rate each item by estimating the percentage of borderline examinees who will answer each item correctly.

1st task: classification of test items into two dimensions; difficulty and relevance

This task could be done by the same panel who will rate each item in the 2nd task, or it could be done by another independent panel

1. Create a two-dimension matrix classifying each item in term of difficulty (easy, medium, difficult) and relevance (essential, important, acceptable, questionable).

2. **Table 10-3.** Example of the two-dimension classification of the test items in the Ebel method.

Items	Relevance / Difficulty	Easy	Medium	Hard
Item 1	Essential			
	Important			
	Acceptable			
Item 2	Essential			
	Important			
	Acceptable			
Item 3	Essential			
	Important			

	Acceptable

3. Regarding difficulty, it could be determined based on the previous data of item – if item already taken from a question bank or used again. The criteria of difficulty based on item analysis are
   - Easy (0.80 – 0.99)
   - Medium (0.45 – 0.79)
   - Hard (0.0 – 0.44)

   If data are not available to categorize item, difficulty could be estimated by the consensus of the panel.
4. Regarding relevance classification, this is should be done the assigned panel for this task. Consensus should be reached to classify each item. If consensus discussion is infeasible, roughly if 50% raters agreed on an item with a two-dimension, it should be considered by this classification.
5. Present the end product of this task as three column table as the following:

Table 10-4. the end-product of item classification in the Ebel method

Items	Difficulty category	Relevance category
Item 1	Easy	Essential
Item 2	Medium	Important
Item 3	Easy	Acceptable
Item 4	Hard	Important
Item 5	Hard	Important

2nd task: rating of test items by the panel

1. Now either the same panel or another independent panel should rate each test item by estimating (minimum pass level, MPL) how many of borderline students will answer this question. The rating form will like the following:

Table 10-5. Rating of test item by the panel in the Ebel method.

Items	Classification		Judge 1	Judge 2	Judge 3	Judge 4	Judge 4
	Difficulty category	Relevance category	45%	50%	60%	45%	55%
Item 1	Easy	Essential	50%	70%	70%	50%	45%

Item 2	Medium	Important	70%	35%	45%	70%	60%
Item 3	Easy	Acceptable	35%	45%	55%	35%	70%
Item 4	Hard	Important	45%	50%	58%	45%	58%
Item 5	Hard	Important	45%	50%	60%	45%	55%

2. Calculate the average of MPL rating for each item

Table 10-6. Calculating the average of MPL

Items	Classification		Judge 1	Judge 2	Judge 3	Judge 4	Judge 4	Average
	Difficulty category	Relevance category	45%	50%	60%	45%	55%	51.00%
Item 1	Easy	Essential	50%	70%	70%	50%	45%	57.00%
Item 2	Medium	Important	70%	35%	45%	70%	60%	56.00%
Item 3	Easy	Acceptable	65%	80%	70%	60%	70%	69.00%
Item 4	Hard	Important	45%	50%	58%	45%	58%	51.20%
Item 5	Hard	Important	45%	50%	60%	45%	55%	51.00%

3. After getting MPL, re-distribution of items based on a two-dimension table

Table 10-7. Re-distribution of items after getting classification and MPL

Relevance / Difficulty	Easy	Medium	Hard
Essential	Item 1 (57%)		
Important		Item 2 (56%)	Item 4 (51.2%) Item 5 (51%)
Acceptable	Item 3 (69%)		

4. If you get more than one item in a cell, calculate the average of MPL for items in that cell

Table 10-8. Making the average of MPL if there are more than one item in a cell

Relevance / Difficulty	Easy	Medium	Hard
Essential	Item 1 (57%)		
Important		Item 2 (56%)	Item 4 + Item 5 (51.1%)
Acceptable	Item 3 (69%)		

5. Covert the percentage into weight and calculate the weighted mean for each raw

Table 10-9. Calculated the weighted means for each raw

Relevance / Difficulty	Easy	Medium	Hard	Weighted mean
Essential	Item 1 (57%)			1*0.57 = 0.57
Important		Item 2 (56%)	Item 4 + Item 5 (51.1%)	1*0.56 + (2*0.51) = 1.58
Acceptable	Item 3 (69%)			1*0.69 = 0.69

6. Calculate the raw passing score by summing of weighted means
   0.57 + 1.58 + 0.69 = 2.85
7. Get the passing score (cut-off point) of the whole test by the percentage of the sum of item weighted means over the number of items
   MPL = 100% * 2.85/5 = 57%

Take-home message

- While developing standards requires considerable effort, it is well worth it when used to certify safe medical doctors.
- Although there are several standard setting methods available, the Angoff method is the most popular and straightforward to use.
- Training judges is the most critical step in any standard setting process.
- There is no a gold standard setting method. For each test and its context, a standard setting method could be appropriate.
- All steps in performing standard setting should be appropriately documented as edvidence to support the validity of decision making of test results.

Further reading

1. Yudkowsky, R., Downing, S. M. & Tekian, A. (2019). Standard setting. In, Assessment in health professions education. Vanderbilt Avenu, New York: Routledge, pp 86-105.
2. Bandaranayake, R. C. (2008). Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Medical Teacher, 30(9-10), 836-845. doi: 10.1080/01421590802402247
3. Ben-David, M. F. (2000). AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher, 22(2), 120-130. doi: 10.1080/01421590078526
4. De Champlain, A. F. (2018). Standard Setting Methods in Medical Education. In: Swanwick, T., Forrest, K. and O'Brien, B. C. (eds.), Understanding Medical Education. West Sussex, UK: Wiley Blackwell, pp 347-359.

References

1. Gasmalla HEE, Tahir ME. The validity argument: Addressing the misconceptions. Medical Teacher. 2021;43(12):1453-5.
2. Wiliam D. Meanings and Consequences in Standard Setting. Assessment in Education: Principles, Policy & Practice. 1996;3(3):287-308.
3. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Medical Teacher. 2008;30(9-10):836-45.
4. Yudkowsky R, Downing SM, Tekian A. Standard setting. Assessment in health professions education. Vanderbilt Avenu, New York: Routledge; 2019. p. 86-105.
5. De Champlain AF. Standard Setting Methods in Medical Education. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding Medical Education. West Sussex, UK: Wiley Blackwell; 2018. p. 347-59.
6. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, editor. Educational measurement. Washington DC: American Council on Education; 1971. p. 508 - 600.
7. Nedelsky L. Absolute grading standards for objective tests. Educational and Psychological Measurement. 1954;14(1):3-19.
8. Ebel R. Essentials of educational measurement Englewood Cliffs, NJ: Prentice-Hall; 1972.
9. Livingston SA, Zieky MJ. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service; 1982.