

Setting a cut score for a performance-based assessment: The Ebel Method

By Dr. Chris Beauchamp

April 21st, 2017

Setting a defensible cut score through a process called standard setting is an essential component that supports exam validity. For traditional multiple choice and other selected response assessments, there is a wide body of literature that supports the use of established standard setting methods such as the Angoff Method (Angoff, 1971) or the Nedelsky Method (Nedelsky, 1954).

For performance-based assessments (PBAs), such as Objective Structured Clinical Examinations (OSCEs), oral assessments or other constructed response assessments, there are also a number of established standard setting methods that can be utilized when there is a relatively large number of candidates. This includes methods such as the Borderline Group Method (Kaufman, et al., 2000) and the Contrasting Group Method (Zieky & Livingston, 1977). However, these methods are not appropriate for PBAs with small candidate volumes. The purpose of this backgrounder is to discuss the use of the Ebel Method (Ebel, 1972) for determining a cut score for a small-scale PBA.

About the Ebel Method

Standard setting methods can be divided into two groups: 1) methods based on judgments about test questions and 2) methods based on judgments about a group of test takers (Livingston & Zieky, 1982).

In methods based on judgments about test questions, such as the Ebel Method, the cut score is established through the concept of the "borderline" candidate. This borderline candidate is one whose knowledge, skills and abilities are on the borderline between being competent or not-yet-competent. The passing score (or cut score) is based on the expected performance of the borderline candidate.

How to implement the Ebel Method

The Ebel Method is comprised of five steps:

1. Select the judges;
2. Define “borderline” knowledge, skills and abilities;
3. Train the judges in the use of the Ebel Method;
4. Collect judgments; and
5. Combine the judgments to determine a passing score.

Step 1. Select the judges: The first step in the Ebel Method is to select the judges. The judges should be able to assess the relative difficulty of assessment content, which generally means that the judges should be practitioners or educators within the field of study. In addition, judges should be representative of the candidate population. This includes representation from different geographical regions and areas of practice. In addition, for entry-to-practice assessments, some of the judges should be relatively new practitioners who are more familiar with entry-level training, experience and requirements.

Step 2. Define “borderline” knowledge, skills and abilities: The second, and possibly most challenging, step is to have the judges arrive at a consensus on the borderline candidate. This involves providing judges with information such as assessment eligibility requirements, candidate demographics and the purpose of the assessment. Judges are encouraged to describe what they feel would be a minimum level of knowledge, skills and abilities. One popular method is to develop performance level descriptors where judges will list characteristics of incompetent, minimally competent and highly competent candidates at each competency category or subject area. This discussion also includes the concept of “forgivable errors” and other characteristics of borderline candidates.

Step 3: Train the judges in the use of the Ebel Method: The Ebel Method is a two-stage procedure. Each judge classifies the questions into groups and then makes a single numerical judgment for each group of questions. The classification of questions into groups is based on two kinds of judgments: A judgment on the difficulty and a judgment on the relevance (or importance). Ebel suggests three difficulty levels; easy, medium and hard; and four relevance categories; essential, important, acceptable and questionable. A judge’s first task is to classify all questions, which results in a classification table. If statistics are available, this information can be made available to judges. The second task is to make judgments about the expected performance of borderline candidates (e.g., “how many borderline candidates will respond correctly to an important/medium question”). Initially, each judge makes one such judgment for each of the 12 blocks of the classification table.

Relevance	Difficulty		
	Easy	Medium	Hard
Essential	4 questions 95% correct	3 questions 85% correct	1 question 80% correct
Important	3 questions 90% correct	3 questions 75% correct	2 questions 60% correct
Acceptable	1 question 80% correct	2 questions 55% correct	2 questions 35% correct
Questionable	1 question 50% correct	0 questions	2 questions 20% correct

Step 4: Collect judgments: The fourth step involves compiling and aggregating the judgments. This is done using the following process:

1. Have the judges make a preliminary classification of the test questions into the 12 categories working individually;
2. Conduct a brief discussion of each question;
3. Have the judges make a preliminary judgment, for each of the 12 categories, of the percentage of such questions a borderline candidate would answer correctly;
4. Conduct a brief discussion for each of the 12 categories and make another round of ratings if required; and
5. Collect judgments.

Step 5: Combine the judgments to determine a passing score: The fifth and final step involves calculating the passing score. This is done using the following process:

1. Multiply the judged percentage correct for the first category by the number of questions (or number of points available) in that category to get the candidate's expected score for the first category;
2. Repeat this step for the other 11 categories; and
3. Add the expected scores for the 12 categories to get the passing score for the whole test.

Relevance	Percent correct	Number of questions	Expected Score
Essential			
Easy	95	5	$.95 \times 5 = 4.75$
Medium	85	3	$.85 \times 3 = 2.55$
Hard	80	1	$.80 \times 1 = 0.80$
<hr/>			
Important			
Easy	90	3	$.90 \times 3 = 2.70$
Medium	75	3	$.75 \times 3 = 2.25$
Hard	60	2	$.60 \times 2 = 1.20$
<hr/>			
Acceptable			
Easy	80	1	$.80 \times 1 = 0.80$
Medium	55	2	$.55 \times 2 = 1.10$
Hard	35	2	$.35 \times 2 = 0.70$
<hr/>			
Questionable			
Easy	50	1	$.50 \times 1 = 0.50$
Medium	*	0	*
Hard	20	2	$.20 \times 2 = 0.40$
<hr/>			
Total		25	17.75

As an additional step, post-administration, the passing rate (percentage of candidates who pass the assessment) based on the passing score should be determined and reviewed for suitability. In the case of results that are unanticipated or problematic for the credentialing organization, small adjustments to the Ebel method passing score can be implemented.

References

Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. In R. L. Thorndike (ed.), Educational Measurement. Washington, DC: American Council on Education.

Ebel, R.L. (1972). Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall.

Kaufman, D. M., Mann, K. V., Muijtjens, A.M.M., & van der Vleuten, C. P. M. (2000). A comparison of standard-setting processes for an OSCE in undergraduate medical education. *Academic Medicine*, 75(3), 267-271.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14(1), 3-19.

Livingston, S. A., Zieky, M. J., and Fowles, M. E. (1978). Manual for setting standards on the basic skills assessment tests. Princeton, NJ: Educational Testing Service.

To read more articles related to eLearning, examination, and instructional design go to www.getyardstick.com and check out our blog.