

Standard Setting Methodologies: Strengths and Weaknesses

John A. Stahl, PhD

Pearson VUE, Chicago, Illinois

Abstract

The purpose of this paper is to present an overview of some of the methodologies used to set criterion-referenced standards on examinations. The paper is divided into two sections. The first section will present those methods that are item/test focused. This section will cover the Angoff, the yes/no, the bookmark, and the item-mapping methods. The second section will cover those methods that are person/work sample focused. This section will cover the body of work, the analytical judgment, and the paper selection methods. The strengths and weaknesses of all of the methods will be discussed.

Standard Setting Methodologies: Strengths and Weaknesses

The determination of a criterion-referenced standard rests on two essential conditions. The first condition is the establishment or presence of a scale of measurement that reflects the underlying construct that is being tested. Tests are created to determine if an examinee has knowledge in a particular field. The tests generally consist of questions/tasks that address specific portions of the more general knowledge area covered by the examination. These questions/tasks are the operational definition of the scale of measurement. Frequently, the questions/tasks are classified according to some pre-established set of knowledge areas, often referred to as the test criteria. This link of the test questions/tasks to criteria is one of the hallmarks of a criterion-referenced test. In some cases, various statistical indices are also attached to the items to position them on the scale of measurement.

The second condition is placement of a point or points on this scale of measurement using a psychometrically sound procedure. These points are frequently referred to as cutpoints. The points demarcate regions on the scale of measurement that are deemed to be different in terms of the purpose of the test. A point may separate a region of pass from a region of fail. Multiple points may separate regions of insufficient mastery from acceptable performance and separate regions of acceptable performance from mastery. An examinee's performance on a test will place them in one of these regions. Since these regions are defined ahead of time, the ultimate classification of an examinee is determined by the operational definition of these regions and not by the examinee's performance as compared to the performance of other examinees. This is a

crucial difference in criterion-referenced examinations as opposed to norm-referenced examinations.

The procedure used to place one or more cutpoints on a scale is commonly called a standard setting procedure or exercise. How these cutpoints are placed on the scale of measurement determine the meaning of the regions thus established. The purpose of the paper is to briefly discuss some of the more popular methods currently used to set standards. Those wishing more information are referred to a recent excellent review of standard setting procedures edited by Gregory Cizek (2001).

The paper will be divided into two sections. The first section will present those methods that are item/test focused. This section will cover the Angoff, the yes/no, the bookmark, and the item-mapping methods. The second section will cover those methods that are person/work-sample focused. This section will cover the body of work, the analytical judgment, and the paired comparison/paper selection methods. The strengths and weaknesses of all of the methods will be discussed.

Item/Test Focused Standard Setting Methods

Item/test focused methods use Subject Matter Expert (SME) judgments on how examinees will perform when presented with items/tasks that comprise a test or an item pool. SMEs are individuals deemed to be knowledgeable in the field being tested and selected to be a representative sample of the practitioners in the field. The SMEs form the panels that provide the information for the determination of the final cutpoint standards.

All standard setting processes begin with a discussion of the expectations for the performance of an examinee who would fall within desired regions on the scale of measurement. In the case of a simple pass/fail decision this involves specifying the

expected performance of a passing individual as opposed to the expected performance of a failing individual. The discussion then narrows down to the expectations for an individual that can just barely pass the desired cutpoint. This individual is frequently referred to as the “Minimally Competent Candidate (MCC).” This hypothetical individual is defined as one who is just qualified to pass the test at a given level (e.g., receive licensure/certification, advance to the next grade).

The Angoff Method

The first procedure for discussion is the Angoff method. In Angoff’s original formulation, SMEs decided whether or not an MCC would answer the item correctly. The cutscore is set as the sum of the correct responses attributed to this hypothetical candidate.

A systematic procedure for deciding on the minimum raw scores for passing or honors might be developed as follows: keeping the hypothetical “minimally acceptable person” in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the scores will equal the raw score earned by the “minimally acceptable person” (Angoff, 1971, pp. 514-515).

In a variation on this method, SMEs estimate the probability of a minimally competent candidate answering each item correctly. The cutscore is set as the sum of those probabilities and is expressed as a percentage. This alternative was discussed in a footnote to the original Angoff paper; however, it has become the predominant Angoff method.

The Angoff method relies on subjective estimations of the ability of an MCC. SMEs base their decisions on the mental image they form of this candidate. In this

standard setting method, a great deal of time is spent refining this concept of an MCC. The expectation of what this candidate can and cannot do is discussed around a framework of the knowledge that must be mastered to obtain a satisfactory score on the assessment.

However, the way that SMEs are asked to use this estimation is problematic. By definition, SMEs are familiar with the content tested on an examination. By the process described above, they refine their concept of the candidates' abilities. The problem arises when they are asked to combine their content expertise and their conceptualization of the ability of the MCC in ways that are outside their expertise.

The Angoff method requires SMEs to make item-by-item decisions. A group of items is selected to represent the content of the examination. SMEs are asked to estimate the percentage of MCCs who will answer each item correctly. The expectation is that a smaller percentage of MCCs will correctly answer hard questions. Conversely, a higher percentage of MCCs will correctly answer easy questions. Being experts in their field, the SMEs have some concept of how hard a particular item will be for a candidate of only minimal competence based on the content of the items. However, the estimation of what percentage of the group of MCCs will correctly answer a question is very difficult for SMEs (Impara & Plake, 1997; Impara, Plake, Hertzog, Giraud, & Spies, 1998). Statistical information on item performance is often given to the SMEs after they provide initial Angoff ratings. Modified Angoff methods provide this statistical information in a variety of forms to the extent that the statistics strongly influence the decisions made by the SMEs. "The performance data . . . provide a reality check so that the expected performance of the BPS is not set unrealistically high or low because a teacher has

misjudged how difficult the item was for the students” (Buckendahl, Smith, Impara & Plake, 2002). SMEs who deviate from the performance data could experience a great deal of pressure to change their ratings and thus question their own judgment.

The Angoff method has several strengths and weaknesses. On the plus side, the SME focus is on the individual item level. It has been the author’s experience in conducting numerous standard setting exercises that SMEs feel comfortable making judgments on individual items relative to a definition of minimal competence. It is also easy to translate the Angoff ratings to a final cutpoint. The SMEs have expressed that they feel more comfortable with not having to make the final cutscore determination. On the negative side, the estimation of the percentage of MCCs that would correctly answer a question requires the SMEs to make judgments that they do not feel comfortable making. (Bejar, 1983; Mitzel, Lewis, Patz, & Green, 2001). SMEs must be trained to perform this task as part of the standard setting exercise, and the degree of success of this training is questionable. Criticism of the Angoff procedure has suggested that it is “fundamentally flawed” (Shepard, Glaser, Linn, & Bohrnstedt, 1993, p. 132).

The Yes/No Method

Impara and Plake (1997) have suggested a return to the original Angoff proposal with a procedure that is called the yes/no method. This method requires the same training of the SMEs in respect to their expectations of the performance of an MCC. It differs however, in the task that the SMEs are asked to perform.

In this method SMEs are asked only to estimate whether an MCC will be able to get an item correct or not. The SMEs go through the set of items individually and indicate their judgment by assigning a yes or a no to each item. The cutpoint or points are

determined by counting the number of yeses. An average across the SMEs becomes the recommended cutpoint.

This procedure is one the SMEs feel very comfortable in performing. They feel more comfortable in saying whether an MCC will get an item right or wrong than in trying to estimate what percentage of a group of MCCs will be successful on the item. The use of statistical feedback is not generally required. The procedure is also relatively easy to implement.

The yes/no method has several strengths and weaknesses. The SMEs feel that making a yes/no decision is much more within their area of expertise rather than estimated a percentage of a particular group.. They can assess what the question is asking and compare that with what they have determined an MCC can or cannot do. The weakness is that the method does not differentiate between how the sum of yeses is derived. For example, on a 50-item test, Judge A could give yeses to the first 25 items and nos to the last 25 items. Judge B could give nos to the first 25 and yeses to the last 25 items. Both judges are setting the same cutscore, but it is clear that they have radically different perceptions of what that cutscore represents. The yes/no method pays little attention to the underlying scale of measurement on which the standard is set.

The Bookmark Method

A method called bookmark has recently become widely used (Karantonis & Sireci, 2006; Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001). In this method, items are ordered by difficulty from easiest to hardest. In the original formulation, SMEs consider the set of items, beginning with the easiest item, and place their bookmark at the item that they believe an MCC would not have a predetermined

probability of answering correctly (usually a 67% probability). Subsequent items in the ordered set are not considered. In recent modifications of the method, SMEs review items beyond the initial bookmark or review the entire set of ordered items. There are generally three ranges identified in the ordered set of items: 1) a region of sure mastery, 2) a region of uncertainty, and 3) a region of sure non-mastery. In this situation the SME places the bookmark within the region of uncertainty.

The bookmark method normally uses Item Response Theory (IRT) calibrations to position the items on a scale of increasing difficulty. It is also possible to use classical test statistics to position the items on a scale; however, this alternative is normally used only for single tests. SMEs are presented with this ordered group of items and are then asked to select a point in the set of items where they feel that an MCC will transition from getting the items correct to getting the items incorrect. Different levels of probability can be designated in setting this transition point. A 50% probability of a correct response or a 67% probability of a correct response are the two most frequent levels used (Lewis, Mitzel, & Green, 1996).

The SMEs review each item and decide if an MCC would answer the question correctly at the specified response probability. The assumption is that the easy items at the beginning of the set of items will be marked as yes and, as the SME progresses to the harder items, there will be a transition point to the items being marked as no. The SMEs are asked to place a bookmark at a transition point where the majority of responses change from yes to no. There are usually multiple rounds of setting the bookmark, with discussions between rounds. Frequently the impact of the proposed cutscores is discussed and, following the discussion, adjustments to the placement of the individual bookmarks

are encouraged (Schultz, Lee, & Mullen, 2005). The final cutscore is the average of the bookmarks.

The bookmark method also has several strengths and weaknesses. Again one strength is that the focus is on item-level data. An additional strength is that the items are arranged in an order that reflects the underlying scale of measurement. This directly ties the standard setting decision to this scale. The SMEs also feel much more comfortable in making yes/no decisions, although the introduction of response probabilities makes these yes/no decisions harder to make. (Impara & Plake, 1997).

On the negative side, SMEs frequently have a great deal of difficulty in placing the bookmark. Therefore, the set of items is often partitioned into three regions; a region of definite incompetence, a region of uncertainty where the MCC would fall, and a region of definite competence.

The size of the area of uncertainty can be quite large, and frequently there is a great deal of variability in the placement of the bookmarks by individual SMEs. Multiple rounds of standard setting are often required to reduce this variability. This process of iteration is designed to reduce the size of the region of uncertainty. At the conclusion of these rounds of standard setting, the mean of the resulting bookmarks is then used to establish the final standard.

Finally, many SMEs feel uncomfortable with making the final determination of where the bookmark should be placed or choosing a cutpoint. They do, however, feel very comfortable in providing information that can be used in making a cutscore decision—that is, how MCCs would be expected to perform on items - but less so when it comes to choosing an actual cutpoint.

The Item-Mapping Method

The item-mapping method is a graphical extension of the bookmarking method. In the bookmarking method, the items are generally presented in the form of a booklet with one item to a page. The items are ordered by increasing difficulty in the booklet. However, it is sometimes difficult to view the items as a set, representing the underlying scale of measurement, rather than just as individual items.

In item mapping, all of the items for a given examination are presented in a histogram form. The items are ordered in columns, with each column in the graph representing a different item difficulty range. The columns of items are ordered from easy to hard, with very easy items toward the left end of the graph and very hard items to the right end of the graph. The difficulties of the items are generally derived from an IRT analysis. An accompanying booklet contains the items represented on the graph.

The SMEs are presented with the graphical representation and the accompanying booklet. The goal of the item-mapping procedure is to identify the column of items on the histogram where judges feel that an MCC has a set probability of answering the items correctly.

The item-mapping procedure shares many of the characteristics of the bookmarking method and thus shares many of its strengths and weaknesses. Again the focus is on the item level, but in the case of the item-mapping procedure the emphasis is more on a set of items represented by a column rather than individual items. The SMEs can also focus more on a region of the examination rather than review the entire set of items. The columns are placed on the scale of measurement so it is easier to tie the cutpoint to this scale. SMEs have also found it easier to make a decision on placing a

cutpoint between two columns of items on the measurement scale than between two items on the scale of measurement.

Person/Work-Sample Focused Standard Setting Methods

In this section, the methods presented use the results of person-performances in setting the desired standards. Person/work-sample focused methods use SME judgments on how examinees performed when presented with items/tasks that comprise a test or an item pool. The judgment and/or standards in these methods are set after the candidates have performed against required items/tasks.

The Body of Work Method

The essence of the body of work method is that an evaluation is based on a collection of a student's work rather than a single sample.

The basis of the body of work (BoW) method is that tests with constructed-response items, there is a better kind of judgment—a kind of judgment for which educators (and others) have greater experience and expertise. The judgment is based on the examination of student responses to a rich body of student work (Kingston, Kahl, Sweeney, & Bay, 2001).

Because the evaluation is based on a body of work, the setting of a standard is much more complex. The general scale of measurement used in a BoW method is normally based on a preexisting set of performance levels and general performance level definitions. In many cases, these levels are mandated by an oversight organization such as a board of education and are frequently established by political bodies.

These general performance-level definitions are then further refined by content specialists, and a list of subject-specific performance-level definitions is developed.

Folders of actual student work are created. The student's work is scored using the scoring for the appropriate test. A variety of folders are prepared representing the range of scores.

A group of SMEs is convened and is thoroughly trained in the meaning of the subject-specific performance-level definitions. The SMEs are then presented with student samples drawn from the appropriate folders and asked to rank them into the desired categories. Initially "range finding" folders are used as a way of gaining a rough approximation of a correspondence between student score and performance classification. Later, "pinpointing" folders are used to refine the process.

The probability of being placed into a particular category is determined from the judgment data and is plotted against the score scale. The point on the resulting curve where that probability is .5 determines the corresponding cutpoint on the score scale.

The body of work method also has several strengths and weaknesses. Its strength is that it can integrate a variety of student work samples into the evaluation process. On the negative side, it requires that an adequate sample of student work be scored prior to the standard setting. The SMEs are required to categorize student work into ranges of proficiency, and the score on the portions of the student's work is frequently presented. This can influence SMEs so that their judgment can be unduly influenced by the revealed scores.

The Analytical Judgment Method

The analytical judgment method (Plake & Hambleton, 2001) is similar to the body of work method in that SMEs are presented with prescored candidate responses. The analytical judgment method tends to be limited to one assessment instrument as opposed to the BoW's rich body of student work samples. The analytical judgment

method also requires pre-established definitions of levels of performance, and the SMEs are trained in the expectations of performance associated with each level.

In the analytical judgment method, SMEs are presented with prescored candidate responses, but the scores are not revealed. As in BoW, the SMEs are asked to sort candidate responses into performance categories, such as advanced mastery, passing, not passing. For each category, SMEs identify a specified number of candidate responses (usually three) that represent the lowest acceptable performance in that category, and they identify a corresponding number of candidate responses that represent the highest performance in the next lowest category. The average score across these identified candidate responses becomes the tentative cutscore. Feedback is provided to the SMEs as to the cutscore, and multiple rounds of the categorization can occur.

The analytical judgment method also has its strengths and weaknesses. The analytical judgment method is simpler to implement than the BoW. They share the requirement that an adequate sample of scored responses be available, however, BoW has a much more extensive sorting task and requires a great deal more preparation time. The initial sorting of candidate responses in the analytical judgment method can be difficult for SMEs. If the quality of candidate responses is relatively homogeneous the sorting can be very difficult and, to a certain extent, arbitrary.

The Paper Selection Method

The paper selection method (Plake, 1998) also uses scored candidate responses. Again pre-established levels of performance are used to set expectations of performance for the desired categories, such as those delineated by the performance of a MCC, and the SMEs are trained in the characteristics of the MCC.

Rather than have the SMEs sort the candidate responses into performance levels, exemplar candidate work is selected for each possible score point. Usually two exemplars are selected for each point. The SMEs are asked to pick the two candidate responses that they feel best represent the work of an MCC. The scores of the candidate responses are not revealed to the SMEs. The average score of the candidate responses selected by the SMEs is the recommended cutscore.

The paper selection method has its strengths and weaknesses. It is one of the easiest candidate-based methods to implement. The process of selecting exemplar papers for each scorepoint can be fairly difficult. It can also be difficult to differentiate papers, particularly in the middle of the score range.

Recommendations and Conclusions

This paper provides a brief overview of some of the standard setting methodologies currently used by test developers. It is by no means an exhaustive review of the methodologies available and is offered only as a sampler of the methodologies available.

The situation that a test developer faces will often dictate what method they will employ in setting standards. In cases in which very little information is available on item performance, such as the development of a new testing program, an Angoff or a yes/no method would be the preferred method. When item performance information is available, a more sophisticated method, such as the bookmark or the item-mapping procedures, would be a better choice.

When constructed response items are used as part of the testing program, the availability of resources will frequently limit the choice in the standard setting

methodology used. In cases of resource constraint, a simpler method such as the paper selection method may be preferred. If more resources are available, a more complex method such as the body of work method may be the method to choose.

Regardless of the procedure selected, a few essential elements of a standard setting exercise should be present. It is essential to remember that most standard setting methods do not result in a single cutpoint. The standard setting methods only yield a defensible range of possible cutpoints. The final decision on a cutpoint is always a policy decision and not one that is dictated by the results of the method used.

The standard setting procedure used should be well documented. Elements of the standard setting procedure that should be documented include:

- The qualifications of the SMEs involved in the procedure.
- The training provided the SMEs in the tasks that they are required to perform.
- The amount of practice provided the SMEs in the standard setting procedures.
- The way in which the standard setting procedure was conducted to include the use of feedback and multiple rounds of standard setting.
- The procedure that was used to determine the final cutpoint/points and the degree of error associated with this determination.
- The information that was provided to the policy organization for the final cutpoint determination.

A standard on a scale of measurement is a crucial element of any testing program and can have a major impact on the assessment of candidates. A well documented and conducted standard setting exercise provides a firm foundation for the application and defense of any cutscore decision.

References

- Angoff, W. H. (1971). Scales, Norms and Equivalent Scores, In R.L. Thorndike (Ed.), *Educational Measurement, 2ed.* American Council on Education. Washington DC.
- Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303-310.
- Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S., (2002), A Comparison of Angoff and Bookmark Standard Setting Methods, *Journal of Educational Measurement, 39:3*, 252-263.
- Cizek, Gregory J. (2001), ed. *Setting Performance Standards: Concepts, Methods, and Perspectives*, Lawrence Erlbaum Associated, Pub. Mahwah, New Jersey
- Impara, J. C. & Plake, B. S. (1997). *Standard Setting: Variations on a Theme by Angoff*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Impara, J. C., Plake, B. S., Hertzog, M., Giraud, G., & Spies, R. (1998). *Utility of a Concept-Focusing Strategy on Judgmental Standard Setting Results*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Karantonis, A., & Sireci, S. G. (2006). The bookmark method: A literature review. *Educational Measurement: Issues and Practice, 25(1)*, 4-12.
- Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L.(2001) Setting Performance Standards Using the Body of Work Method, in *Setting Performance Standards: Concepts, Methods, and Perspectives*, Gregory J. Cizek ed., Lawrence Erlbaum Associated, Pub. Mahwah, New Jersey
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard Setting: A Bookmark Approach, In D.R. Green (Chair) *IRT-based procedures using behavioral anchoring*. Symposium conducted at the Council of Chief School Officers National Conference on Large Scale Assessment, Phoenix, AZ.
- Lunz, M. E., Stahl, J. S., & Wright, B. D. (1994). Interjudge Reliability and Decision Reproducibility. *Educational and Psychological Measurement, 54*, 913-925.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark Procedure: Psychological Perspectives. In G.J. Cizek (Ed.), *Setting Performance Standards*. Mahwah, NJ.
- Plake, B.S., (1998) Setting performance standards for professional licensure and certification: Implications for National Assessment of Educational Progress, *Applied Measurement in Education, 11*, 65-80.

- Plake, B.S., & Hambleton, R.K., (2001), The Analytical Judgment Method for Setting Standards on Complex Performance Assessments, in *Setting Performance Standards: Concepts, Methods, and Perspectives*, Gregory J. Cizek ed., Lawrence Erlbaum Associated, Pub. Mahwah, New Jersey
- Schultz, E. M., Lee, W., & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement*, 42, 1-26.
- Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G., (1993), *Setting performance standards for student achievement: A report of the National Academy of Education Panel of the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels*. Stanford CA: Stanford University, National Academy of Education.