test DEVELOPMENT

Fundamentals for Certification and Evaluation

MELISSA FEIN

chapter 11:

Standard Setting

Key objectives covered in this chapter:

- Define standard setting and recall major synonyms for minimum passing score.
- Visualize and describe a minimally competent person and the role this person plays in the standard setting process.
- List the major components of the standard setting process and the documentation that should be collected and retained.
- Define and differentiate the major standard setting methods, including the bookmark procedure, Angoff variations, contrasting groups, and borderline group. Describe the link between rubrics development and standard setting.
- Identify the role of standard setting in program evaluation contexts.
- Conceptually describe what the standard error of measurement (SEM) is and why it is important.

Myth: 70 percent is passing.

Introduction

Standard setting, in the context of criterion-referenced test (CRT) development, describes the process used to determine the thresholds of minimally acceptable performance levels. When there are two categories of performance, it is used to differentiate between those who do have mastery of a given subject matter or skill set versus those who do not. The desired outcome of most standard setting sessions is a fair and defensible *minimum passing score*, also called a *cutoff score* or *cut score*. If there are more than two categories of performance, such as less than proficient, proficient, and advanced, then standard setting involves establishing multiple cutoff scores. This chapter covers the fundamental process issues involved in standard setting. Two standard setting methods are discussed in depth: a modification of the bookmark procedure suitable for use with the classical test theory (CTT) approach to test development and a variation on the Angoff method. Two additional standard setting methods, contrasting groups and borderline group, are summarized. Since cut scores are not perfect, recommendations and caveats related to reporting measurement error are reviewed. The use of standard setting in rubric development is briefly addressed. In addition, the role of standard setting in program evaluation contexts is discussed.

Process Issues

The main process issues involved in standard setting revolve around establishing a standard setting team, facilitating the standard setting team in working together to create a vision of a minimally competent person, motivating the team members to agree to personally take the assessment, managing standard setting activities so that closure is made on a defensible decision within the time constraints faced by the team, and managing team member absences and personnel changes within the team.

A standard setting team should be comprised of subject matter experts (SMEs), a planner, and a facilitator. All participants on the standard setting team need to sign confidentiality agreements, including the planner and facilitator. The SMEs are the team members who are making the standard setting decision, with the guidance of a facilitator; the planner handles logistics related to the standard setting activities. Regardless of the specific method used to set standards, the credentials of the SMEs need to be documented as a part of the standard setting process. This provides evidence supporting content validity. The credentials that need to be documented include relevant positions held; the amount of time served in those positions; relevant education, training, licenses, registrations, certifications and certificates held; general contact information; and references. The planner should have some general understanding of the standard setting process but does not need to be an expert. He should also understand the organization's procurement regulations. And although it may sound superficial, the planner should understand the feasibility, and appropriate scale and style of the standard setting event being planned—in terms of details related to travel, accommodations, and refreshment options during the sessions; expectations and standards of appropriateness can vary greatly from organization to organization. He should also have the ability to communicate effectively with critical decision-making personnel within the organization. The facilitator should have expertise in the standard setting process and be able to manage the group dynamics of a standard setting session in person or electronically. The planner and facilitator functions can be accomplished by the same person, but they do not have to be. Sometimes it is optimal for the planner to be a member of the organization engaged in standard setting and the facilitator to be an outside consultant. Or both positions can be effectively hired out. If neither position is hired out, however, it is critical that the in-house employee who is responsible for facilitating the sessions has adequate expertise as well as the independence to do the job. In other words, the facilitator should not be put in the position of having to run a session that is being directed by SMEs who have an agenda in

terms of the outcome, or who might be that facilitator's supervisor now or in the near future. From an accountability perspective, this is a reason to consider hiring out the facilitation component of the work.

The first order of business in any standard setting session, after giving team members a general orientation regarding the purpose of the session and expectations, is to have the SMEs define a *hypothetical minimally competent person.* There is almost always a humorous backlash to this request. The SMEs perceive that the whole reason they are involved in exam development is to ensure that the highest standards are being met, and they often find it ironic that they are being asked to think in terms of minimal competence. The issue is that most SMEs have optimal or ideal competence foremost in their minds and they need to make a shift from the idea of optimal performance to envisioning the lowest performance level that a person can have and still be considered competent. Facilitating a productive discussion on the concept of minimal competence can enable the SMEs to (hopefully) come to a shared vision of what this means operationally, in terms of applying this to standard setting activities. Success of the standard setting session depends on some type of consensus, agreement, or at least a manageable degree of disagreement on this concept.

It is valuable to have the SMEs personally take the exam as if they were examinees. Even though items have (hopefully) been adequately reviewed prior to the standard setting session, additional item flaws or scoring problems can become startlingly obvious when the exam is actually administered to SMEs on the standard setting team. Some SMEs will not be happy about being asked to take the exam; they may suggest that they are the experts and should not be required to submit to this activity, and others will worry that they will do poorly and be embarrassed. SMEs can be reassured. Scores should not be shared. Each SME will know her score, and this will inform her participation in the standard setting session. Team members can be reminded that they are members of the standard setting team because they are the experts. They are not being judged;

they are the judges and their insight on the standard setting activities will be more fully informed if they experience the items from the perspective of an examinee.

Effective management of the standard setting session requires that the session planner be aware of who is participating, why they are participating (including potentially hidden agendas), and any constraints related to time, interest, and effort. In workplace contexts, SMEs are sometimes drafted by supervisors or colleagues to be a part of a test development and standard setting team. The task is sometimes perceived as an add-on to an already full set of job responsibilities. In credentialing contexts, SMEs might be volunteers who also serve in some leadership capacity within the credentialing organization, or even serve in a position on its board of directors. These team members are generally people who are strongly committed to the organization and are volunteering their time. Often, team members may be geographically dispersed, making in-person standard setting sessions a challenge to schedule. Because of these factors, it is particularly crucial for standard setting sessions to be well planned and orchestrated in a manner that recognizes and values the above-and-beyond time and effort investment of those involved. In addition to conventional courtesies and expressions of gratitude, team members often feel less put upon when the planner shares the rationale underlying the need for standard setting activities that consume significant amounts of effort and time.

Management of standard setting sessions will vary depending on the method used, the session format (for example, in-person versus electronic), and the assessment format. In-person standard setting sessions are the most desirable format for many reasons, one being that when teleconferencing, nonverbal communication can be missed in coming to a shared vision of minimal competency, which is a core theme of the session. However, web conference sessions, which provide screen sharing or video, are feasible in many contexts. The reality of the resources available for standard setting activities cannot be ignored, but

lack of resources is not an excuse to cut corners. If an in-person standard setting session is not going well, and team members find themselves spinning their wheels, the facilitator should urge them not to make a rushed decision for the sake of having closure, but should get agreement on the feasibility of having some type of follow-up session, either in-person or via a teleconference or web conference.

Ideally, members of the standard setting team will be present for all sessions and none will drop out and need to be replaced. Realistically, people will miss sessions and a member or even several members might need to be replaced. To minimize this occurrence, the planner should recruit team members who understand the time commitment and can be counted on to fulfill their responsibilities. When attrition happens (in spite of optimal recruitment strategies), the planner will need to organize a process to ensure that those who have missed meetings are caught up on decisions as well as the rationale supporting those decisions. This might take the form of providing the absentee member with standard setting session minutes, along with a one-on-one training summary of what transpired. When members are replaced, it is particularly important to ensure that new members are trained to understand the process and rationale that went into the shared vision of the hypothetical minimally competent person.

Bookmark Procedure

The bookmark procedure is a relatively new, more streamlined standard setting method designed to reduce the cognitive effort and time burden on the SMEs. Although it was developed for use in the item response theory (IRT) approach to test development, the modification presented here is suitable for use with the classical test theory approach to CRT development, which is the focus of this book. The bookmark procedure requires pilot data, so if these data are not available, another standard

setting method must be used. Based on the pilot data, the item difficulty for each item is computed. Then a newly ordered version of the exam is constructed in which items are ordered by difficulty, from easiest to hardest. This ordered exam is the version of the exam administered to the SMEs. The idea behind this method is that the SMEs can work through the exam and find the point of difficulty at which the hypothetical minimally competent examinee might not be expected to answer correctly. The sum of the number of items correctly answered up to this point would define the cutoff between passing and failing (Mitzel et al., 2001).

A standard setting session using the bookmark process includes the following steps:

- SMEs discuss the knowledge, skills, and abilities (KSAs) and the expected performance level of a hypothetical minimally competent person.
- The reordered exam based on pilot data with items ordered from easiest to most difficult is administered to the SMEs. A key is provided at the end of the exam, for self-scoring. See the technical appendix for computational instructions related to item difficulty (chapter 16).
- Each SME takes the exam, and makes a personal judgment identifying the item at which he believes the cut score should be set. SMEs are also instructed to make notes on any items they see as problematic, for any reason.
- The mean, variance, and range of each SME's cut scores are computed. See the technical appendix for computational instructions (chapter 15).
- SMEs discuss these cut scores in the context of the hypothetical minimally competent person and the KSAs, including:
    - The implication of selecting specific cut scores, such as

the pass/fail rates that would be expected. Impact data such as pass/fail rates should not drive the decision, but can be informative. For example, if only 10 percent of the examinees appear to be passing and if many who failed are known personally by the SMEs and are believed to be competent, this might suggest the cut score is too stringent.

- Consequences of decisions, such as the risks or costs to society or the organization of allowing incompetent people to be credentialed, as well as the consequences of failing examinees who are actually competent, should be reviewed and deliberated.
- A discussion of the role of the exam versus other indicators of mastery in the credentialing process would also be an appropriate part of the discussion. Credentialing decisions are not supposed to be based solely on exam results; there should be other components or indicators of competency.

- SMEs are asked to individually revisit the cut score based on the group discussion.
- SMEs share and discuss any changes in their preferred cut score.
- The above discussion process iterates until a decision is made, either a decision on the final cut score or a decision to defer the decision, reconvene, and finish at a later time.
- A summary of the process, discussions, and decision(s) should be documented.

It should be noted that the bookmark method can be used with a variety of item formats. In addition, it lends itself to establishing multiple mastery categories, such as proficient and advanced, with minimal additional cognitive burden; standard setting team members would simply be placing multiple bookmarks in the ordered exam instead of one.

Angoff Variations

The idea behind the Angoff method is that SMEs consider how the hypothetical minimally competent examinee might be expected to perform on each item. There are a number of variations on the Angoff method. In one variation, which is suitable for use with a multiple-choice item format, SMEs go through the exam and award an item score of "1" if they believe a minimally competent examinee can be expected to get the item correct, and a "0" otherwise. The cut score is comprised of the sum of the item scores (Angoff, 1971; Cizek et al., 2004; Cizek, 2006; Impara and Plake, 1997).

A standard setting session using the Angoff variation described above includes the following steps:

- SMEs discuss the knowledge, skills, and abilities (KSAs) and performance level of the hypothetical minimally competent person.
- SMEs take the exam individually, are provided with a key, and denote for each item whether a minimally competent person would be expected to get the item correct, summing up the expected number correct and recording that as the cut score.
- SMEs discuss each item in the context of the minimally competent person. If pilot data on item difficulty are available they should be included for use in this discussion and SMEs should compare the item difficulty data to their own opinions. Instructions for computing item difficulty are presented in the technical appendix (chapter 16).
- The mean, variance, and range of the cut scores are discussed along with implications of selecting specific cut scores, as described in the bookmark method (i.e., impact data, consequences, and so on).

- The above process iterates until a decision is made, either a decision on the final cut score or a decision to reconvene and finish at a later time.
- A summary of the process, discussions, and outcome should be documented.

It should be noted that in other variations on the Angoff method for multiple-choice items, SMEs are asked to come up with a numerical probability (falling between zero and one) that a minimally competent examinee would answer an item correctly or, alternatively, they are asked to indicate the proportion of minimally competent examinees who would be expected to answer each item correctly.

When the item format is constructed or extended response, in which item scores can take on a range of values rather than being simply correct or incorrect, as with multiple-choice items, the Angoff modification is referred to as the *extended Angoff variation*. For the extended Angoff, the SMEs designate for each item the number of points they expect that a minimally competent person might score. As in the method described above, these item points are summed over all of the items to come up with a cutoff score.

It was noted that the bookmark modification must be done with pilot data, but the Angoff variations may be done with or without pilot data. There are arguments both for and against having standard setting sessions with and without pilot data. These arguments will not be reviewed here. There is no definitive rule about this and the decision is usually driven by contextual and pragmatic issues.

Additional Standard Setting Methods: Contrasting Groups and Borderline Group

Two additional standard setting methods that can be used with assessments comprised of any item format are the contrasting groups method and the borderline group method. In contrast to the bookmark procedure and the Angoff method, in which the focus is on the test items, in the contrasting groups and borderline group methods the focus is on identifying people who perform at a particular competency level.

In the *contrasting groups method*, the goal is to administer the exam to known masters and non-masters and then look at the difference in the score patterns for those two groups. The cut score is defined by the point at which the score distributions for the different groups intersect (Crocker and Algina, 1986).

In the *borderline group method*, the goal is to administer the exam to individuals having borderline competency. After the exam is administered to a group of individuals with borderline competency, the cut score is defined as the median score of these borderline examinees (Crocker and Algina, 1986). Instructions for computing a median are provided in the technical appendix (chapter 15).

Accurate identification and selection of groups is crucial to the success of the contrasting groups and borderline group methods. In some cases, members of the standard setting team might not be in a position to identify and recruit masters, non-masters, or individuals with borderline competency. It is not sufficient or appropriate for the standard setting team to ask random supervisors to conscript examinees based on perceived competency, unless those supervisors have a working understanding of the agreed upon vision of the hypothetical minimally competent person, provided through some type of orientation or training. The role of the standard setting team in the contrasting groups and the borderline

group methods is to discuss and agree upon a vision of the hypothetical minimally competent person and come up with an operational definition that can be communicated to and applied by SMEs who are in a position to identify and select people of the groups. It should be noted that a major challenge in using the contrasting groups and borderline group methods, aside from the selecting the group members accurately, is acquiring a large enough pool of examinees. In a small scale testing program with few expected examinees, these methods might not be feasible to use because the estimates will not be stable or representative when computed on such a small group of examinees.

Standard Setting in Rubric Development

A *rubric* is a scoring tool used to organize the grading of an item which has a structured, extended response that is worth a variable number of points, depending on the quality of the response. This is unlike a multiple-choice item which is either completely correct or completely incorrect. Because rubrics involve the assignment of numerical scores to observable work behaviors and work product quality, SMEs need to be involved in integrating or linking these point values to the vision of the hypothetical minimally competent person. If the extended Angoff method of standard setting is used, the point values of each item will be examined by the SMEs in the context of the expected performance of a minimally competent person as they go through the items. If other standard setting methods are used, point values associated with rubrics used in item scoring still need to be justified through a standard setting process at some point, and the process needs to be documented.

Standard Setting and Program Evaluation

When criterion-referenced tests are used in training program evaluation contexts, standard setting may or may not be relevant. If the focus of a program evaluation is to look at the effectiveness of the training program in terms of ensuring that participant competency levels are met through training, then standard setting is relevant because competency needs to be defined to see if it is being met. The focus of some training program evaluations, however, is on determining program effectiveness in terms of impact on participant improvement rather than a fixed level of competency. In these program evaluation contexts, the gain in knowledge, defined by the difference between post-test and pretest scores, is of primary interest. In this situation a program might be considered effective if post-test scores of participants are significantly, in a statistical sense, higher than pretest scores. Alternatively, some program funders want effectiveness to be defined by a specific percentage gain in post-test scores over pretest scores. There is nothing to prevent a program evaluation from addressing both gain and absolute competency. What generally occurs is a reflection of what the program funder wants to see.

Imperfection of Cut Scores and the Standard Error of Measurement

Cut scores are imperfect. This section gives an example of imperfection in individual scores as a vehicle for conceptually understanding the *standard error of measurement* (SEM). Then the use of the SEM is extended to the context of *imperfect cut scores*. Recommendations and caveats for the use of the SEM are highlighted. Hopefully the reader will not confuse SEM, standard error of measurement, with SME, subject matter experts. It is regrettable that two such similar acronyms needed to appear in the same chapter.

Even the best tests are imperfect measuring devices; there is error associated with any score that is reported. Imagine an examinee, Carla, who takes an exam once and earns a score of 87. If she takes the same exam again five more times, it is probably unlikely that she will get a score of exactly 87 on each of those subsequent tries—assume she gets equivalent items and cannot study between retakes. Suppose she gets the following scores: 88, 86, 87, 89, 85, 86. Looking at all of her scores, one realizes that her real competency is not best represented by pinpointing her score at 87, but maybe it is better represented by indicating that it falls within the range between 85 and 89. Informally we might say that Carla's score is an 87, plus or minus two points. That two-point buffer takes error into account and it can be thought of as an *error band*. In this example, the highest and lowest scores are used as the boundaries on the error band. In real CRT development contexts, the computation of the error band is more complex, but only slightly, and that error band is referred to as the standard error of measurement (SEM).

Cut scores selected by a standard setting team are also subject to error, just as individual scores are. For this reason it is recommended that the SEM be reported for cut scores. It should be noted that this recommendation has a complication for those developing small scale testing programs. Score reliability or precision varies with examinee competency and the best practices recommendation is to compute the SEM near the cut score. If there are multiple cut scores, a SEM should be computed near each one. However, computation and interpretation of this might be problematic if the number of examinees is small. Instructions for computing the SEM are provided in the technical appendix (chapter 18). The SEM should be computed and entered into the test development documentation. If the number of examinees is very small, it should be interpreted with caution.

Summary

The purpose of standard setting is to provide a fair and defensible process for determining who is granted a particular level of mastery status and who is not. The detailed activities of the SMEs on the standard setting team vary depending on the standard setting method used, but what is common to all methods is that the SMEs create a common vision of the hypothetical minimally competent person. In the bookmark procedure and the Angoff method variations, the focus is on considering the items, or content in the context of the expected performance of that hypothetical minimally competent person. In the contrasting group and borderline group method, the focus is on finding a pilot examinee group of real people who reflect performance above and below—for contrasting groups—and at the performance level of that hypothetical minimally competent person for borderline group. The standard setting process should be practical to implement, documented, and the result should be replicable if team members are swapped out or replaced. The development of scoring rubrics includes the documentation of the linkage between points awarded in the context of KSAs and expected performance. In some program evaluation activities, conventional standard setting is irrelevant. Instead, program effectiveness is determined by looking at the degree of improvement over performance measured prior to training, rather than considering an absolute competency standard. Even cut scores set with the most diligent of standard setting teams are subject to error. Computing the SEM on a cut score provides an indication of scope of the error. All standard setting activities should be documented, as should the qualifications of those on the standard setting team.

References and Resources

Angoff, W.H. (1971). Scales, Norms and Equivalent Scores. In *Educational Measurement*, ed. R.L. Thorndike (pp. 506-600). Washington, D.C.: American Council on Education.

Cizek, G.J. (2001). *Setting Performance Standards: Concepts, Methods and Perspectives*. Mawah, NJ: Lawrence Erlbaum Associates, Publishers.

Cizek, G.J. (2006). Standard Setting. In *Handbook of Test Development*, eds. S.M. Downing and T.M. Haladyma (p. 225-258). Mawah, NJ: Lawrence Erlbaum Associates, Publishers.

Cizek, G., M. Bunch, and H. Koons. (2004). An NCME Instructional Module on Setting Performance Standards: Contemporary Methods. *Educational Measurement: Issues and Practice*, 23, 31-50.

Crocker, L., and J. Algina. (1986). *Introduction to Classical & Modern Test Theory*. Orlando: Harcourt Brace Jovanovich College Publishers.

Harvill, L.M. (1991). An NCME Instructional Module on Standard Error of Measurement, www.ncme.org/pubs/items/16.pdf http://www.mendeley.com/research/ncme-instructional-module-standard-error-measurement/ (accessed October 11, 2011).

Impara, J.C., and B.S. Plake. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, 35, (1), 69-81.

Mitzel, H.C, D.M. Lewis, R.J. Patz, and D. Green. (2001). The Bookmark Procedure: Psychological Perspectives. In *Setting Performance Standards: Concepts, Methods and Perspectives*, ed. G.J. Cizek (pp. 249-281). Mawah, NJ: Lawrence Erlbaum Associates, Publishers.

Related Topics in the Technical Appendix

- SEM: Computation (chapter 18)
- Ordering Items by Difficulty (chapter 21)
- Median Computation: Borderline Groups (chapter 21)
- SME Credentials Form (chapter 21)
- Standard Setting Facilitator's Worksheet (chapter 21)
- Confidentiality Agreement Components (chapter 21)

The page intentionally left blank.