Promoting best practices for the credentialing community

ICE RESEARCH & DEVELOPMENT COMMITTEE

Standard Setting Overview for Credentialing Programs

Lawrence J. Fabrey, PhD (Chair)
Andrew C. Dwyer, PhD
Chuck Friedman, PhD
Jerry B. Reid, PhD
Patricia Young, MA

Institute for Credentialing Excellence™

Published by

Institute for Credentialing Excellence
2025 M Street, NW, Suite 800
Washington, DC 20036

Printed in the United States of America

2018 Research & Development Committee

Chair: John Wickett, PhD, Wickett Measurement Systems
Vice-Chair: Jerry B. Reid, PhD, American Registry of Radiologic Technologists

Members:

Kirk Becker, PhD, Pearson VUE
Susan Davis-Becker, PhD, ACS Ventures LLC
Andrew Dwyer, PhD, American Board of Pediatrics
Lawrence Fabrey, PhD, PSI Services LLC
JG Federman, EdD, Creighton University
Chuck Friedman, PhD, PSI Services
Michelle Gross, MBA, MEd, NCC, NCSC, LPC, National Board for Certified Counselors
Peg Harrison, MS, CPNP-PC, CAE, Pediatric Nursing Certification Board, Inc.
Kari Hodge, PhD, NACE International
James Magruder, Board of Registered Polysomnographic Technologists
Patricia Muenzen, MA, ACT ProExam
Lisa Nepi, EdM, SeaCrest Consulting
Nicole Risk, PhD, American Medical Technologists
Greg Sadesky, PhD, Yardstick Assessment Strategies
John Weiner, MA, PSI Services LLC
Daniel Wilson, Mountain Measurement, Inc.
Patricia Young, MA, Kryterion, Inc.
Anthony Zara, PhD, Pearson VUE

Thank you to Kristin P. Walinski for her professional editorial services and to ICE staff member Doug Weinbaum for his support of the Research and Development Committee.

Thank you to our 2018 Research and Development contributors!

Dennis Faulk Ambassador

National Board for Certification in Occupational Therapy
National Board of Certification and Re-certification for Nurse Anesthetists
National Commission for Health Education Credentialing Inc.

Leader

Global Skills Exchange Corporation
National Athletic Trainers' Association Board of Certification, Inc.

Patron

Agilutions Consulting
American Association of Medial Assistants
Construction Manager Certification Institute
National Council of Architectural Registration Boards
Oncology Nursing Certification Corporation
Prometric
PSI Services

Partner

APICS, Inc.
Board of Pharmacy Specialties
Denise Roosendaal, CAE
Financial Planning Standards Council
HR Certification Institute

Table of Contents

Disclaimer: *The Institute for Credentialing Excellence (ICE) is the professional organization that provides education, networking, research, and other resources for credentialing professionals. ICE's research may include evaluation of emerging practices in the credentialing community, coverage of practices that credentialing organizations have undertaken in response to challenging situations, or promotion of exemplary best practices. Every product of the R&D committee represents the consensus judgment of the individuals on the respective task force and is reviewed by the committee as a whole. ICE's research products are not necessarily intended to be a guide for obtaining accreditation or for complying with the Standards offered through ICE: the National Commission for Certifying Agencies (NCCA) Standards for the Accreditation of Certification Programs, the ISO/IEC 17024: Conformity assessment programs, or the ICE 1100 Standard for Assessment-based Certificate Programs*.

Introduction

This paper provides an introductory overview of standard-setting processes, methods, approaches, challenges, issues, and policies for credentialing organizations.[1]

Standard setting for credentialing programs typically involves a judgment-based process for identifying the examination score that reflects the minimum level of qualifications required to earn the credential: that is, which candidates meet the criterion versus which candidates do not. In this context, the word *qualifications* could be defined differently by different organizations, with common variations including *competence*, *knowledge*, or some other term. Sometimes standard setting is considered the overarching intent of determining qualifications, whereas establishing a cut score can refer to the operational process used to provide information for those responsible for setting the standard.

In presenting key issues faced by credentialing organizations, this paper is organized into the following major sections:

1. Overview of Standard Setting: Brief Definition and Purpose
2. General Components of Standard Setting
3. Standard-Setting Methods
4. Policy Decisions: Selecting the Standard
5. Standard-Setting Literature
6. Current Areas for Further Discussion
7. References

The focus of this paper is on multiple-choice examinations, but many of the basic principles can be applied with modification to performance-based and oral examinations. This paper provides a structure of key points for credentialing organizations; it does not replace the hundreds of articles and books on standard setting or the knowledge provided by psychometricians**.**

Overview of Standard Setting: Brief Definition and Purpose

Standard setting is an attempt to identify how much is enough. In the context of a credentialing examination, the “how much” may be said to refer to knowledge, skill, performance, or proficiency. This level is typically translated into a dichotomous decision, that is, either pass or fail, which will lead to a decision either to certify or not to certify the individual.

First, from the National Commission for Certifying Agencies (NCCA) *Standards for the Accreditation of Certification Programs* (National Commission for Certifying Agencies, 2014), standard setting is defined as:

> A systematic method for determining the passing score on an examination based on the characteristics of the examination, particularly its level of difficulty. The result of the process is a pass/fail cut score that represents the lowest level of acceptable performance in the content area being assessed by an examination.

[1] Different organizations may define the process as “establishing a passing point” or “setting a cut score”; this paper will generally use *standard setting*. In addition, this paper will usually use the word *credentialing* to refer to both certification and licensure. However, in some instances, only certification is used, even when the concept could also apply to licensing.

The second definition is drawn from the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) glossary, which defines standard setting as "the process, often judgment based, of setting cut scores using a structured procedure that seeks to map test scores into two discrete performance levels that are usually specified by performance-level descriptors."

It is difficult to separate the definition of standard setting from its purpose, as evidenced by these two definitions. Measurement involves the process of assigning numbers to quantities of something, which could be associated with distance, time, height, or weight, for example. With credentialing examinations, numbers may be assigned to represent knowledge, skill, performance, or proficiency, resulting in some score continuum. The purpose of standard setting is to find a point along the score continuum that reasonably balances the purpose for which examination scores will be used with the limitations inherent in all examinations.

There may be a controversy about whether there is a true standard for an examination. Some involved with credentialing may believe that a true standard exists for an examination, and the goal of a study is to find that standard. Others may believe that there is no true standard, and the goal is to identify a standard that provides evidence about the level of competence that will need to be demonstrated to optimize various policy issues. The reality of any scale of measurement is that by dichotomizing the results to classify examinees into one of two mutually exclusive categories (certified or not certified), a great deal of information differentiating individuals along a continuum (i.e., the range of "scores") may be discarded.

For example, if we are counting correct answers (i.e., with classical test theory), one point will differentiate passing or failing. Psychometricians sometimes say that standard setting is arbitrary but not capricious. It is arbitrary in the sense that the process relies on judgments that cannot be proved to be correct. However, the process is not capricious in that there is a rational basis for the judgments.

Standard setting is an operationalization of the exam's statement of purpose that should be linked to the credentialing program's statement of purpose. The process of going from the mission to the end result of credentialing involves a complex sequence of translating the high-level policies to operational definitions. For example, a certification program that is intended to identify individuals competent to function in a particular role must operationalize that role through a job analysis that delineates what a person in the profession is expected to be able to do.[2] The role is used to identify the knowledge, skills, abilities, and other characteristics (KSAOs) that an individual must have mastered to competently perform the role. Individuals' degrees of mastery of the KSAOs are distributed along a continuum from a low degree to a high degree.

Test items are developed that assess an individual's mastery of the construct of interest, and the items are assembled into an examination with appropriate proportions of different knowledge areas that mirror the examination specifications as derived from a job analysis. The examination results in a score that places each candidate on the continuum of mastery. Standard setting selects a point on the continuum allowing the credentialing decision of sorting individuals into two groups: those who "have enough" and those who "don't have enough."

Clearly, selecting the standard has important consequences for the credentialing program since it relates through the chain of decisions back to the mission statement of the credentialing organization. Even if all other steps in the operationalization process are conducted with the greatest attention to alignment back

[2] A job analysis study is sometimes known by other terms, such as a *practice analysis* or *role delineation study*.

to the organization’s mission, a cut score that results in classification decisions of individuals that seem inconsistent with the examination’s purpose can undermine the quality of the credentialing process.

General Components of Standard Setting

This section provides an overview of the major components and considerations required in any standard-setting process, regardless of the method or approach. It addresses considerations for collecting data, choosing participants in the data-collection process, and defining the minimally qualified candidate (MQC)**.**

Data-Collection Process

Perhaps the most salient component of standard setting is the data-collection process, which provides a rational basis for the judgments. However, turning the judgments of subject-matter experts (SMEs) into numbers and producing various statistics from those numbers should not cloud the fact that ultimately selecting a standard is a policy decision, not a calculation. If standard setting were simply a matter of estimating the “correct” pass/fail value, then a policy decision might not be needed. However, one standard cannot be shown to be more “correct” than another standard in the typical situation since an external criterion is not typically available for such a determination. Calculations alone cannot determine what standard best meets the purpose of the examination.

Participants in the Data-Collection Process

SMEs

The central component of standard setting involves the collection of professional judgments from SMEs. As such, the panel of SMEs should be selected in a way that minimizes the potential for unwanted bias. At the individual level, each SME should be sufficiently knowledgeable with respect to both examination content and the target population. Most often, this is achieved by selecting SMEs who are already certified, but other noncertified individuals, such as those who supervise certificants on the job, may also have this critical knowledge and experience.

In addition to considering the individual qualifications of each SME, there are important reasons to pay attention to the panel’s group-level characteristics. Large-group activities and discussions are typically part of the earlier stages of a standard-setting study, and these discussions and activities affect the judgments made by the individual SMEs in the later stages of the study. Also, it is common practice for the recommended performance standard from the entire panel to be obtained by aggregating the individual SME recommendations (each SME weighted equally). For both of those reasons, it is important to structure the panel so that all important subgroups are adequately, but not overly, represented.

To accomplish the desired level of balance, the panel should be selected to reflect the certificant population with respect to key demographic variables such as gender, race, age, geographic location, and practice/workplace setting. In addition, it is usually important to include at least a few recently certified individuals on the panel, as they often have more realistic and accurate expectations with respect to the MQCs than those who are further removed from the certification process. The inclusion of educators or training providers on the panel deserves special consideration due to the potential for conflicting interests, especially in cases where pass rates on the credentialing examination are used to (formally or informally) evaluate the educators or trainers.

Buckendahl and Davis-Becker (2012) discuss panel composition in depth, including a detailed discussion regarding the potential conflicts of interest that individuals belonging to various subgroups may bring to the panel. They suggest that the panel should, wherever possible, consist of three groups (in addition to ensuring the panel is demographically representative):

- recently credentialed practitioners,
- experienced practitioners, and
- educators

Some caution should be exercised related to the inclusion of these groups on a panel. The NCCA accreditation standards provide guidance about panel composition. For example, while faculty members may appropriately serve on a panel, an educator who provides preparatory materials for the credentialing examination would not. Similarly, overemphasis from one group of practitioners could lead to a potential conflict of interest, if that group were inclined to "set the bar too high."

Psychometrician to Guide the Process

One key participant in standard-setting studies is the psychometrician, or testing expert, who oversees the process and leads or facilitates the standard-setting activities. This testing expert is expected to provide training for the SMEs to ensure they understand three things: (1) the overall process, (2) the role all individuals are expected to play within that process, and (3) the specific assignments the SMEs are being asked to complete. In addition, the testing expert must be able to identify and thwart potential threats that may emerge and undermine the validity of the final performance standard. For example, several standard-setting activities within the process may involve large-group discussion (e.g., defining a profile for the MQC). During group discussion, the testing expert must make sure that each panelist has a voice and that no single panelist dominates the discussion in a way that would unduly influence the overall panel's final decisions or recommendations.

Testing Sponsor Representative

It may be beneficial for a representative from the credentialing organization (e.g., the certification program director or member of the governing board) to participate in certain standard-setting activities, especially large-group discussions where information pertaining to the credentialing program would be useful (e.g., establishing a profile for the MQC). For example, this representative may be able to provide insight into the organization's mission or historical information about the examination or information about previous standard-setting studies. In special cases, this representative may also have the content expertise to serve as a standard-setting panelist. Not only would this representative be able to provide useful information to the SMEs during large-group discussions, but this individual may also serve as a liaison to the governing board when it is tasked with reviewing the panel's recommendation and making a final policy decision regarding the passing standard. As with the SMEs, the testing expert leading the standard-setting study should take steps to ensure that this representative does not influence the panel in ways that would undermine the validity of the final performance standard.

Defining the Expectations of a Minimally Qualified Candidate

As mentioned in the opening section of this paper, standard setting for credentialing programs typically involves a judgment-based process for identifying the score on the examination that reflects the minimum level of competence required to earn the credential. In short, standard setting involves a panel of experts who are asked to determine the score an MQC would likely earn on the examination. A highly critical component of any standard-setting process, therefore, is the identification of the relevant knowledge, skills, abilities, and other characteristics (KSAOs) possessed by an individual who is at least "minimally qualified" for the credential.

While there is no universally accepted methodology for creating an MQC profile, several useful exercises can be applied in most credentialing settings.

Reviewing eligibility criteria: Reviewing the criteria that must be met to sit for the examination is often a first step in creating an MQC profile because it helps define the minimum qualifications of the overall test-taking population. For example, if a credentialing program requires candidates to have earned a specific degree from an accredited training program and to have acquired at least 3 years of full-time experience on the job, it would follow that the MQC would possess, at a minimum, the KSAOs associated with those requirements. In other words, a credentialing program's eligibility requirements often are a starting place for creating a useful MQC profile.

Defining profiles for unqualified and highly qualified candidates: One inherent assumption underlying the credentialing process is that the candidate population consists of individuals whose knowledge and skills range from incompetent (and unworthy of the credential) to highly competent, with "minimal competence" falling somewhere between those two extremes. Because standard setting focuses on the MQC, identifying the characteristics that distinguish the MQC from the unqualified candidate (who still meets eligibility requirements) and that distinguish the MQC from the highly qualified candidate can be useful. In other words, defining separate profiles for unqualified, minimally qualified, and highly qualified candidates may help to provide a clearer definition of the MQC.

Identifying job tasks that can be performed independently: Another exercise that can be used to define the MQC is to identify the critical job tasks (as well as the knowledge and skills needed to perform those tasks) that are expected to be performed independently (i.e., without supervision) by a newly certified individual. Also, identifying the job tasks that are either not performed by newly certified individuals or not entrusted to newly certified individuals without close supervision can help distinguish the MQC from individuals with a more advanced skill set. Similarly, an exercise can be conducted in which the SMEs identify tasks that would be expected to be uniformly easy or uniformly difficult for the MQC, to help set the context for the discussion of item ratings.

Standard-Setting Methods

This section of the paper briefly describes a few of the most commonly used standard-setting methods employed by credentialing programs. All the methods described in this section rely on the expert judgments of SMEs who possess knowledge of both examination content and the target population. These SMEs are asked to review an examination form (or a set of exam items) and to judge how an MQC would perform on each item or one of the collective set of items. The descriptions below attempt to explain the (sometimes subtle) differences between these methods and to highlight the specific advantages or disadvantages they may provide. A much more thorough description of each of these methods, including the psychometric theory underlying them, can be found in *Setting Performance Standards: Issues and Practice* (Cizek, 2012).

For any standard-setting method that involves SMEs, it is essential that the SMEs sign a nondisclosure or confidentiality agreement before seeing any of the test questions. The agreement should clearly define the terms of the agreement (e.g., not to discuss the examination questions, examination content, or confidential aspects of the examination-development process; not to retain copies of the examination questions or other confidential material, or not to compromise the security of the examination materials while in your possession) and require the SME's signature. The agreement should also include a conflict of interest clause that details activities that would be considered a conflict of interest with examination-development activities (e.g., participating in third-party examination preparation or delivery services, providing private instructions to potential examinees) and the designated period that the SME must abstain from these activities. The SMEs must also not be eligible for taking the examination, whether for certification or recertification purposes. The nondisclosure or confidentiality agreement and conflict of interest agreement are required to meet Standards 10 and 11 of the 2014 NCCA Standards and to protect the security of the examination materials.

Modified Angoff Method

According to the ICE Handbook (Institute for Credentialing Excellence, 2009), the modified Angoff method was the most commonly used method for setting performance standards on credentialing exams in 2008, and there is little reason to believe that has changed in recent years. Because there are so many variations of this method, however, it is perhaps more accurate to say that the most commonly used standard-setting method comprises the family of Angoff methods. Within the Angoff family, all methods involve SMEs who provide a separate judgment for each item. The original Angoff method asks "each judge to state the probability that a 'minimally acceptable person' would answer each item correctly" (Angoff, 1971, p. 515). A common variation of this method, the "Yes/No method" (Impara & Plake, 1997), requires SMEs to answer a slightly simpler question, "Would an MQC answer this item correctly?", which requires only a yes (correct) or a no (incorrect) instead of estimating the probability of a successful response.

The Angoff methods vary from each other with respect to other factors as well. For example, in many standard-setting activities, SMEs who have provided judgments for each item are then offered some feedback and given an opportunity to revise their earlier judgments based on that feedback. The type of feedback provided can vary (e.g., item difficulty information or expected pass rate information), along with the number of opportunities SMEs are given to revise previous judgments (often two or three rounds of judgments).

Once the final judgments of the SMEs have been recorded, the judgments are averaged across items and across SMEs. This statistical estimate represents the score an MQC would be expected to achieve on the examination, with lower scores reflecting less than minimal competence and higher scores reflecting more than minimal competence. This score estimate is presented to the governing or decision-making body as the recommended passing standard. The narrative below continues an explanation of the Angoff process.

How to train and calibrate SMEs?

For the modified Angoff method (and many other criterion-referenced methods), it is essential that SMEs are trained on the process as well as oriented on the MQC profile described earlier. The training for a modified Angoff process typically includes a review of the following:

- the purpose of the activity,
- the steps in the process,
- what Angoff ratings mean and how to make them,
- how Angoff ratings translate into a passing score,
- the considerations in making Angoff ratings, (e.g., risk of not having the knowledge being assessed, item phrasing, and difficulty of distractors), and
- the logistics of the process.

It may be helpful to include a discussion that cut scores for credentialing examinations cannot be put on the same scale as academic examinations (e.g., a cut score of 65% is a "D"); the difficulty of the items must correlate to the expected performance of the MQC for that credential.

A calibration exercise should be included in the modified Angoff process training to ensure that SMEs have a similar understanding of the expectations for an MQC. The calibration exercise involves having the SMEs make Angoff ratings on a few items (perhaps between three and ten items) until the SMEs feel comfortable with the process and they are producing Angoff ratings within an acceptable range (e.g., a 20- or 30-point spread). During the calibration exercise, it can be helpful to have additional discussion among the SMEs on their rationales for their ratings for illustration purposes.

How often and when to share item difficulty statistics?

A measure of item difficulty (e.g., *p*-value or item response theory (IRT) *b* value) may be provided to the SMEs after the initial set of ratings for some or all ratings. These measures are frequently used as a reality check to SME participants, especially when the group of SMEs has rated the difficulty of the item for the MQC as significantly different from how the item has historically performed. It is not a requirement that the average Angoff values for items be highly correlated with the item difficulties, but it indicates that the SME's judgment of item difficulty for the MQC relates to actual item performance. It may be best to withhold item difficulty feedback if candidate counts are not sufficiently large for reliable item statistics or if the candidate sample is not representative of the candidate population.

There is no consensus in the field regarding the practice of providing item difficulty statistics to the SMEs during the modified Angoff process. In some variations of the modified Angoff process, item difficulty statistics are provided on all items, while in others, they are provided only on certain items (e.g., when the average Angoff rating differs significantly from the item difficulty statistic). Some research can inform the decision regarding how and when to use item difficulty statistics. For instance, Busch and Jaeger (1990) found that when *p*-values were shared, there was a higher correlation between Angoff ratings and *p*-values. There was also a higher inter-rater reliability when *p*-values were shared before the final iteration of Angoff ratings. Norcini, Shea, and Kanya (1988) studied the degree of influence of sharing normative data (e.g., *p*-values). They found that normative data influenced raters about 25% of the time, but raters had a low average change per item. Raters changed their ratings mostly on items whose difficulty they rated as very high or very low.

Which items to gather Angoff ratings on: scored only or also beta test?

The decision regarding which items to include in the modified Angoff process relies on practical considerations. If a program is gathering Angoff ratings on items or an examination that has not been beta tested (i.e., it does not know which items will be used as scored items), it is necessary to include all items that will appear on the beta examination (or are eligible to be used as scored items on an exam). If a program knows which items will be used as scored items on the examination, the Angoff ratings process can be limited to only the scored items.

If there is more than one form of the examination and a sufficiently large number of candidates, the scored items for one form can be included in the standard setting and the cut score for the second form determined via an equating study. If a program has low candidate counts, there may not be a sufficient amount of data to determine the cut score of new forms using an equating study. In these cases, programs may be forced to rely on Angoff data to set the cut score for the new or revised test form. This approach may not be consistent with the commentary of the NCCA accreditation standards. Equating, new programs, and low volume are discussed in the Current Areas for Further Discussion section of this paper.

How many rounds of ratings?
In the modified Angoff process, a minimum of two rounds of ratings is recommended. The first round frequently involves SMEs making independent ratings on the items. If the modified Angoff process involves only two rounds, the second round involves discussing items on which the first-round ratings are too varied and modifying ratings is desired. Common reasons for modifying Angoff ratings include that the SME was not as familiar with the content measured by the item or was swayed by another SME's rationale. It may be possible to combine the first two rounds by seeking independent ratings of the first item, discussing it and allowing SMEs to change or record a second rating, and then moving on to all subsequent items.

During the first round of Angoff ratings (i.e., initial ratings), the SMEs may or may not be given access to the correct answers to the items. There may be some benefit to not providing the correct answers to the items because SMEs may rate the item as easier for the MQC when they have the correct answer and they are not forced to look at the item as a candidate would.

One variation of the modified Angoff process requires SMEs to take the test before making any Angoff ratings. The rationale behind this variation is to provide a reality check and orientation of the test's content, breadth, and rigor. However, taking the examination first can make SMEs overly familiar with test items, which makes them seem easier when assigning Angoff ratings.

Other variations on the modified Angoff method ask SMEs to re-rate every item to ensure that they did not skip an item they intended to re-rate or adjust. These additional rounds introduce some additional type of feedback (e.g., item difficulty values or the correct answers to items).

As noted in a previous section, a third round of ratings may occur, with SMEs being provided with item difficulty statistics. And, as noted earlier, this third round could be incorporated as SMEs provide the ratings. In other words, with the three rounds combined, SMEs could:

- read the item,
- provide a rating,
- select a response (or confirm the key as provided),
- discuss the item with other judges,
- record a second rating (or change the initial rating),
- review the item difficulty statistics (i.e., *p*-value or b parameter estimate),
- record a third rating (or provide a single, final rating), and
- repeat this sequence for every item on the examination

Bookmark Method

The Bookmark method is perhaps the second most prevalent standard-setting method used in practice behind the Angoff method. Overall, its prevalence can largely be attributed to its usefulness for establishing multiple performance standards within a single assessment to distinguish between multiple performance levels. For example, many educational accountability assessments attempt to classify students into performance groups (e.g., basic, proficient, and advanced). In cases like these, the Bookmark method is preferred because using the Angoff method to establish three separate performance standards on the same assessment would be overly burdensome. However, most credentialing examinations require only one performance standard; thus, the Bookmark method has been slow to gain widespread use for credentialing examinations.

Like the Angoff method, the Bookmark method requires a panel of SMEs to judge how the MQC would perform on the examination. More specifically, the items within the examination are ranked according to their difficulty level and presented to the SMEs in that order. Each SME is then asked to identify the point (i.e., the "bookmark") that separates the set of items that the MQC would be expected to answer correctly from the set of items that would be expected to be answered incorrectly. For each SME, this point can be interpreted as his or her recommended passing standard. The average of these recommendations across all SMEs is presented as the panel's recommended passing standard.

While the basic premise underlying the Bookmark method is relatively straightforward, the implementation of this method requires practitioners to make several nuanced policy and psychometric decisions. For example, in the previous paragraph, it was stated that each SME would be asked to identify the set of items that a "MQC *would be expected* to answer correctly." In practice, there are several possibilities for operationalizing what the phrase "*would be expected*" means. In some cases, that phrase may simply be interpreted as the point at which an MQC would have at least a 50% chance of answering that item and all easier items correctly (referred to as the 50% response probability, or RP50, in the standard-setting literature). Others have advocated, especially in educational standard settings, for using a response probability criterion of 67% (RP67), meaning that SMEs would be asked to identify the set of items that the MQC would have at least a 67% chance of answering correctly.

In addition to selecting a response probability criterion, the Bookmark method also requires all items to have statistical information that can be used to order the items according to their difficulty level. Most often, items are ordered by difficulty estimates that have been computed under an IRT model, which allows items from separate administrations of the examination to be placed on the same item difficulty scale. IRT item statistics also allow items to be ordered according to different response probabilities (e.g., RP50 or RP67). But it is possible to conceptualize the Bookmark method from a classical test theory perspective as well.[3]

Borderline Group Method and Contrasting Groups Method

The Angoff and Bookmark methods (see previous sections) can be labeled test-centered standard-setting methods because they require SMEs to judge how a hypothetical candidate (i.e., the MQC) would perform on the test. Examinee-centered standard-setting methods represent a different class of methods that involve obtaining a group of real examinees, identifying each examinee's level of competence using an external criterion (i.e., other than the examination), and then investigating their performance on the examination. Two examples of examinee-centered methods are the borderline group method and the contrasting groups method.

In the borderline group method, "borderline" candidates are identified as being just barely qualified for the credential based on an external criterion. These borderline candidates are also asked to take the examination, and their performance on the examination (e.g., average score) is used as the passing

[3] Specific guidance regarding the implementation of the Bookmark method is beyond the scope of this paper. More detailed information regarding the Bookmark method can be found in Lewis, Mitzel, Mercado, and Schultz (2012).

standard recommendation. In the contrasting groups method, candidates are classified into groups (e.g., qualified and unqualified) based on an external criterion. Subsequently, exam score distributions for both groups are examined, and the point on the score scale that minimizes the number of misclassified candidates (e.g., “qualified” candidates who would fail and “unqualified” candidates who would pass) is treated as the passing standard recommendation.

Clearly, the major challenge for examinee-centered methods is identifying an external criterion that can be used to classify candidates based on their level of competence. In many cases, no such external criterion exists, and even in situations where a valid external criterion may exist, it is often costly to obtain external criterion scores/classifications for a sufficiently large number of candidates. As a result, examinee-centered methods are rarely used in practice.

Supplemental Methods: Hofstee and Beuk

Some programs opt to use supplemental standard-setting methods, such as the Hofstee (1983) or Beuk (1984) methods. These methods are sometimes used in conjunction with another standard-setting method (e.g., modified Angoff or Bookmark). They both operate to consider or compare a normative with an absolute standard (i.e., percentage of candidates who pass with appropriate cut score for the examination).

The Hoftstee method (more accurately, a modification to the Hofstee that is commonly used) involves asking SMEs for the upper and lower limits of the percentage of the candidates who would pass the examination as well as the upper and lower limits of the acceptable passing scores. The average of the recommendations is applied to a cumulative percent distribution of test-taker scores. The cumulative percent distribution of scores should contain enough candidates to be stable, and the candidates should be representative of the target audience for the credential.

The Beuk method involves asking SMEs to provide the minimum score that candidates should possess to pass the examination (expressed as a percentage score) and the percentage of candidates that the SMEs would expect to pass with that cut score. The means and standard deviations of these percentages are calculated. The point represented by the two means (mean of the minimum score predictions, mean of percentage passing rate predictions) is plotted on a graph of the actual passing rate at each score. A line with a slope equal to the standard deviation of the passing percentage predictions divided by the standard deviation of the minimum score predictions is drawn through this point to intersect the curve of the actual passing rates at each score. Using the Beuk method, the passing score is the point at which this line intersects the curve.

Other Factors Considered in Selecting a Recommended Cut Score

After going through a judgment-based process such as the modified Angoff methodology, many credentialing programs find it useful to consider other information to help inform or validate the final cut score decision. Kane (2001) suggested considering procedural, internal, and external information when evaluating the cut score recommendations from a standard-setting process.

- Procedural evidence includes who participated in the process, the appropriateness of the standard-setting methodology, SME participant feedback on the implementation of the process (e.g., Did they think the process was fair and produced a fair outcome? Did they think they were too lenient or too harsh in their judgments in relation to the MQC?), and the measurement characteristics of the examination.

- Internal evidence includes consistency of SME participant ratings (e.g., correlation between an individual SME's ratings and item difficulty values) or the convergence of SME ratings (e.g., level of agreement or standard error of mean or median).

- External evidence includes results of secondary standard-setting methods (e.g., Hofstee) and historical pass rates for the examination. Licensure programs might also consider practice-based disciplinary action in the occupation to ensure the passing standard mitigates the risk to the public.

Policy Decisions: Selecting the Standard

Who Selects the Standard

While it is possible to consider standard setting in terms of the procedural steps followed, thinking about standard setting within the overall context of the credentialing process can be instructive. How one thinks about standard setting has implications for how it is performed, who participates, and what their roles are. If standard setting is considered a straightforward data collection and statistical calculation, then a group of SMEs directed by an individual with expertise in the standard-setting process may be the primary participants with the results presented to the policymaking body (e.g., certification board) as a fait accompli. If, on the other hand, standard setting is considered a more interactive policy-formulation process, the certification board may be more intimately involved in the process.

There are proponents for each of these approaches. Those who advocate for minimal board involvement point out that including the certification board in the process may result in board members having too much influence, thus decreasing the value of input from an independent group of SMEs. Those who advocate for more board involvement point out that as the group with fiduciary duty for the credentialing program, they should be fully engaged in all stages of standard setting and be knowledgeable about the process since they will be responsible for selecting, and perhaps, defending the standard. The process of setting a cut score on the examination is fundamentally about clarifying values and expectations as it relates to the construct being assessed and further back to the organization's mission.

The most appropriate level of certification board involvement will depend upon the particular circumstances. For example, if there are individuals on the certification board who are not SMEs, they will not have the level of knowledge needed to effectively participate in evaluating individual items under the Angoff procedure. However, deriving the description of an MQC requires a different type of knowledge that even the non-SME may possess and may be an appropriate activity. If the certification board members decide to limit their involvement in the details of the process, the two critical policy decisions that they should be encouraged to be engaged in are defining the MQC and selecting the actual cut score based upon the information that the SMEs generate.

Adjusting a Recommended Cut Score

At the beginning of this paper, the process of identifying a recommended standard was said to be arbitrary, but not capricious. Similarly, adjustments by the governing body to the recommended cut score from the standard-setting study may be arbitrary but should not be capricious. In other words, adjustments may be made to the recommended cut score, but the adjustment should be based on a clear rationale. One approach taken by governing bodies is to recognize that any type of measurement (quantification by assigning a number) is subject to statistical error. These are not errors in the sense of “mistakes” but rather errors in the sense of variability inherent in estimating anything using a sample versus the population. There are techniques for estimating the amount of variability or error in various processes. Use of these as a basis for adjustments to a recommended cut score is possible.

Consider two specific types of error:

- Scores produced by examinations are subject to measurement error. For example, an examinee taking the same examination multiple times would be expected to achieve slightly different scores. Also, the same examinee taking different forms of an examination measuring the same construct would be expected to obtain slightly different scores. If an examinee were to take the examination many times or take many versions of the examination, a distribution of scores for that examinee would result. The standard error of measurement can be used to estimate the deviation of a particular score from the mean score for that distribution of scores.

- A cut score produced by a panel of SMEs would also be expected to vary if the process were replicated across multiple occasions or replicated with a different panel of SMEs. Again, a distribution of cut scores would result. A standard error can be used to estimate the deviation of the particular recommended cut score from the mean cut score for that distribution of recommended cut scores.

Discussing the implications of using various indicators of error should be discussed with a psychometrician so that the governing body uses a clear rationale when considering an adjustment to the recommended cut score.

What Evidence Is Used to "Validate" the Standard?

Standard 17 of the NCCA Standards for the Accreditation of Certification Programs addresses standard setting and suggests factors that should be addressed in the report documenting the standard-setting process. This information should be made available to the board. These factors may include the following:

- the method used and why it was selected,
- the number and qualifications of the judges,
- any deviations from the process (including judges who provide data that suggests a lack of understanding of the process),
- the definition of minimal competence used (depending upon the standard-setting method used),
- the potential risk of errors in classification (i.e., weighing the risk of passing unqualified candidates vs. failing qualified candidates), and
- the impact on passing rate.

Determining a cut score for an examination is not just a result of data analysis. Historical trends, some fallibility of the judges and the process, new test configurations, and policy considerations must be factored in as well.

> Standard-setting studies utilizing procedures such as the Bookmark or Angoff methods are just one component of the complete standard-setting process. Decision makers ultimately must determine what they believe to be the most appropriate standard or cut score to use, employing the input of the standard-setting panelists as one piece of information among multiple sources. (Geisinger & McCormick, 2010)

If the board views standard setting as a psychometric calculation, it may be inclined to simply accept the cut score generated from the standard-setting study as "correct." If the board understands that standard setting is a structured approach to assist the decision makers to make an informed decision regarding the score that optimizes multiple (and sometimes competing) factors, then it will better realize its role in selecting the cut score. Since the board is the final arbiter on what cut score will be used, it should be fully educated on the nature of standard setting.

Standard-Setting Literature

With respect to standard setting, perhaps no other current text provides a more thorough explanation of the theories underlying this process and a more comprehensive review of the various standard-setting methods employed than Setting Performance Standards: Theory and Practice (Cizek, 2012). This book also includes a chapter devoted to the important issues that should be considered when setting a performance standard in a credentialing context (Buckendahl & Davis-Becker, 2012).

Other helpful resources that provide an overview of standard-setting methodology include the ICE Handbook, the Handbook of Test Development, Testing in the Professions (in particular, chapter 6 on interpreting scores), and Educational Measurement: Issues and Practice (numerous articles on standard setting). The ICE Handbook is of special interest because it also provides information about the prevalence of various standard-setting methods that NCCA-accredited certification programs are using. As testing professionals might expect, the modified Angoff approach (or some variation of that approach) was identified as the most commonly used method for setting a performance standard on a credentialing examination. Despite rapid changes in technology, which allow for different types of assessments and provide unique opportunities for collecting standard-setting data from panels of SMEs, there is little reason to believe that the modified Angoff approach has been supplanted by another standard-setting method as the method of choice for credentialing programs.

Current Areas for Further Discussion

Frequency of Standard-Setting Studies

The NCCA *Standards for the Accreditation of Certification Programs* provide good guidance in the commentary for standards related to job analysis and standard setting:

> Standard 14 commentary: “As a general guideline, a job analysis should be conducted every five years. However, for fast-changing professions, occupations, roles, or specialty areas, an analysis every one to three years may be more appropriate. Similarly, when content is not expected to change rapidly, certification programs may find it appropriate to wait as long as seven to eight years between job analyses.”
>
> Standard 17 commentary: “A standard-setting study should be conducted following completion of each job analysis study at a minimum but can be conducted more frequently to support programmatic requirements.”

Therefore, it can be concluded that certification organizations will generally reassess their examination passing standard about every five years. Some “triggering events” that could lead to an organization reconsidering a standard might relate to changing expectations in relation to the certification program. Any stakeholder group, such as consumers, employers, applicants, or the policymaking body itself, may initiate these expectations. While a criterion-referenced standard-setting method is intended to identify a cut score irrespective of the percentage of candidates who pass, if the examination cut score is no longer serving the program’s intended purpose, it may be time to re-evaluate the cut score.

Normative Considerations

The rationale behind using a criterion-referenced passing standard (as opposed to a norm-referenced one) is well understood in the credentialing industry. If the passing standard were norm-referenced, then a candidate's certification status would depend on that candidate's knowledge and skill relative to his or her peers within a specific test administration window (norm-referenced) as opposed to depending on whether the candidate's knowledge and skills are sufficient for safe and effective performance on the job (criterion-referenced).

It is ironic, then, that despite an awareness of the importance of using a criterion-referenced passing standard, many credentialing programs continue to use pass rate information (i.e., norm-referenced information) as the primary source of evidence for validating the passing standard. To give an extreme example, if a criterion-referenced passing standard produced a pass rate of either 0% or 100%, it is almost certain that the credentialing organization would adjust the passing standard (or conduct a new standard-setting study) to produce a more reasonable pass rate. In fact, it is common practice, although not universally endorsed by testing experts, for passing rate information to be included as part of the standard-setting process. For example, after making an initial round of ratings or forming an initial cut score recommendation, panelists may be presented with "impact data" (i.e., "here is what the passing rate would be for the panel's recommended cut score") before being given an opportunity to modify their ratings and/or their final passing standard recommendation. In addition, the decision-making body may want to consider the passing rates for all the potential cut scores within the recommended range when making their final decision.

While it is safe to say that most, if not all, psychometricians understand the importance of using a criterion-referenced approach to standard setting, it is also clear that in practice, standard setting is rarely, if ever, successful in removing all normative considerations. And, in some cases, normative considerations may improve the validity of the passing standard (and, by extension, the credential itself). Regardless, there is clearly a need for continued conversation regarding the role that normative data should play in establishing and/or validating the passing standard.

Equating vs. Standard Setting

One potential misuse of standard-setting methodology is the practice of setting a new passing standard on each newly developed form of an examination. There may be situations in which that practice may be acceptable, but generally speaking, standard setting is the process for establishing a criterion-referenced passing standard for a single form, while test equating is a statistical procedure specifically designed to ensure that the passing standard established for one form is maintained across several forms (usually forms in subsequent administrations or years). Because test equating is a statistical procedure that requires examinee-response data, there may be situations where test equating is either not possible (i.e., no response data exists for the new form at the time that the passing standard needs to be set) or where the equating estimates lack sufficient precision (i.e., often with small or nonrepresentative samples of examinees).

Several equating methods have been proposed over the past decade that have been specifically designed for small samples (e.g., Circle Arc and Nominal Weights Mean). In addition, a recent study (Dwyer, 2016) found that equating with examinee samples as small as 25 was more effective for maintaining an equivalent passing standard across forms than conducting a new standard-setting study.

It is beyond the scope of this paper to identify specific situations in which conducting a new standard-setting study on each newly developed form would be sound psychometric practice. The important point is that standard-setting methodology was not designed to maintain an equivalent passing standard across forms. Thus, in situations where test equating is impossible or overly problematic, certification programs are encouraged to explore other options for maintaining an equivalent passing standard across forms, including identity equating or modifying the standard-setting methodology in such a way as to maximize the likelihood of holding all candidates to the same passing standard across forms.

New Certification Program

It should be clear from this paper when a criterion-referenced passing point study should be conducted, but logistical questions could arise in relation to the study for a new certification program. As noted previously, the SMEs for standard setting should be sufficiently knowledgeable with respect to both examination content and the target population, and this is most often achieved by selecting SMEs who are already certified. With a new program, this is not possible, so the certification organization must find other ways to evaluate the qualifications related to knowledge of examination content. Generally, this is not an issue, as most certification organizations will develop and apply policies and procedures related to the composition of the groups of SMEs that are necessary to create and maintain a certification examination. Since NCCA *Standards for the Accreditation of Certification Programs* Standard 8 provides that “certification program must award certification only after the knowledge and/or skill of the individual candidate has been evaluated and determined to be acceptable,” it raises the question whether the standard-setting panel should be granted the credential. This is a fairly common practice, and certification organizations that opt to award the credential use different methods to ensure that the knowledge of panel members has been evaluated and deemed acceptable. One method may involve only a review of the education, experience, and other qualifications of the panel members. A stronger method is to compute a score for the panel members during the standard-setting study and require the panelists to provide a response when recording the initial rating.

Low Candidate Counts

The issue of low candidate counts may not often be thought of as point of discussion related to standard setting. Part of the issue has already been discussed, as a previous section has suggested that equating is nearly always preferable to repeating a standard-setting study to maintain a comparable standard. The other issue may be that with a low candidate volume for a new certification program, it is sometimes difficult to evaluate the reasonableness of the result. While there are undoubtedly exceptions, it is common for the initial candidate group to be more experienced and of relatively higher ability than subsequent groups. If the initial group is small and the passing rate is high, that may be because of the higher ability, but that may not be the case. The issue becomes more significant when trying to maintain the same standard of competence.

In-Person vs. Remote Standard-Setting Meetings

It is becoming more common for credentialing organizations to conduct standard setting via a series of remote meetings rather than through a single in-person meeting. There are advantages and disadvantages to both in-person and remote meetings, some of which will be discussed here.

A major disadvantage of in-person meetings is the potential expense for the organization, that is, travel for participants and expenses associated with the venue (e.g., meeting room rental and catering). In addition to the higher cost of in-person meetings, another potential disadvantage is that participants may work at different paces, which can create dead time during the meeting for some and cause others to feel rushed to complete the activity.

The focus of a standard-setting study is to seek consensus on the definition and interpretation of the MQC. Face-to-face meetings are thought to be generally effective at accomplishing this objective. In-person meetings allow the credentialing organization to ensure full participation and gain consensus from the SMEs and to better monitor the security of the test items. If a certification organization has a tradition of using in-person meetings for standard setting, the organization should consider these issues before deciding to change to remote meetings for standard setting.

A primary benefit of remote meetings is the potential cost savings associated with not incurring travel expenses. Another benefit of remote meetings is that it may be easier to schedule participants for a few remote meetings than an in-person meeting, and the organization may get participation from SMEs for whom it is difficult to travel. In addition, remote meetings may allow an organization to make more effective use of the SME's time, in that when the activity is completed, the SMEs can simply leave the meeting rather than "waiting for their flights." If a remote meeting can provide a process that allows for the development of necessary panel consensus, this methodology can be advantageous.

A common concern related to remote meetings is a perceived examination security risk because SMEs have to be given access to operational test items without direct supervision. There are systems or means for securely giving item access to SMEs, but the credentialing organization is trusting that they will not take copies of these questions or be careless when accessing the test items (e.g., leaving the computer unattended when viewing the test items). Another potential disadvantage of remote meetings is that all SMEs may not participate in all phases or meetings due to unexpected circumstances. Standard-setting processes often require the same set of SMEs to participate in all phases and meetings, and this can be problematic for having a sufficient number of SMEs to maintain the group's representativeness of the credential's target audience. A general criticism of remote meetings is that participants may not remain fully engaged in the meeting since they are attending from their work or home environments with all the demands that can arise.

Alternative Item Types

Because an overwhelming majority of credentialing examinations consist of traditional multiple-choice items, it is not surprising that most of the standard-setting literature explicitly or implicitly focuses on that item type. As measurement science and technology continue to evolve, however, an increasing number of credentialing organizations are exploring alternative item types, defined broadly as any item type that is not multiple choice, to improve the measurement qualities of their examinations. For those looking for more information on alternative or innovative item types, the ICE and the Association of Test Publishers (ATP) recently published a white paper that discusses innovative item types in detail, including what alternative types of items are being used, the rationale behind choosing to use those item types, their measurement benefits, and the challenges associated with their use (Institute for Credentialing Excellence, 2017).

With regard to standard setting, most of the important principles for setting a standard on a multiple-choice examination also hold for an examination that includes alternative item types. For example, the Angoff question asked of SMEs for each item on a multiple-choice examination is, "What percentage of minimally qualified candidates would answer this item correctly?" That question implies, at the least, that each item has a single correct answer. For item types with more than one correct answer or items that can be scored using a partial credit scoring model, the Angoff question may need to be slightly rephrased as, “How would a minimally qualified candidate score on this item or on the examination as a whole?” Although some of the methodological details may depend on the specific item type, the overarching principles of defining minimal competence and obtaining judgments regarding how an MQC would perform on the examination are universal across all item types.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.), Washington, DC: American Council on Education.

Beuk, L. S. (1984). A method for reaching compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, *21*, 147-152.

Buckendahl, C., & Davis-Becker, S. (2012). Setting passing standards for credentialing programs. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, innovations*. New York, NY: Routledge.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, *27*, 145-163.

Cizek, G. J. (Ed.) (2012). *Setting performance standards: Foundations, methods, innovations.* New York, NY: Routledge.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement*, *53*, 3-22.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers*. Educational Measurement: Issues and Practice*, *29*, 38–44.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Hemlick (Eds.), *On educational testing* (pp. 109-127). San Francisco, CA: Jossey-Bass.

Institute for Credentialing Excellence. (2017). *Innovative item types.* Collaborative white paper of the ATP Beyond MC Items Committee and the ICE Innovative Item Type Task Force. Washington, DC: Institute for Credentialing Excellence.

Institute for Credentialing Excellence. (2009). Certification: *The ICE handbook*. Washington, DC: Institute for Credentialing Excellence.

Impara, J., & Plake, B. (1997). Standard setting: an alternative approach. *Journal of Educational Measurement*, *34*, 353-366.

Kane, M. (2001). So much remains the same: Conception and status of validation in standard setting standards. As cited in G. Z. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

Lewis, D., Mitzel, H. C., Mercado R. L., & Schultz, M. (2012). The Bookmark standard setting procedure. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, innovations*. New York, NY: Routledge.

National Commission for Certifying Agencies. (2014). *Standards for the Accreditation of Certification Programs.* Washington, DC: Institute for Credentialing Excellence.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement*, *25*, 57-65.