



General Considerations for Setting a Passing Standard ¹

Assembling a Subject Matter Expert Panel

In order for examinations to be scored, a process known as *standard setting* is performed by a Subject Matter Expert (SME) panel. The panel members are assembled to represent the diversity of practice, education, experience, training, age, gender, ethnicity, and geographic practice settings that characterizes the profession. This diversity and expertise provides the foundation for the successful establishment of the standard. Each member of the panel team brings a unique perspective, be it as a highly experienced practitioner nearing retirement, to one who has recently achieved certification or a license. Standard setting relies upon empirical knowledge – personal, experience based understanding – of the profession. Individual SMEs work to achieve consensus by melding together these diverse and rich experiences and opinions, providing support for the decisions made collectively.

Considering the Entry-Level Practitioner

The first exercise in establishing a passing standard for an examination is to review or develop a profile of **candidate eligibility**. In order to qualify for credential candidacy, eligibility requirements have been established that candidates must meet before they are allowed to begin the examination process. These requirements may include education, apprenticeship, minimal age requirement, on-the-job experience, internship or documentation of self-study. Program-specific eligibility requirements will be discussed to establish a profile of candidates meeting eligibility to begin the exam process.

The purpose of the certification/licensure program is to identify candidates possessing an established level of **minimal competency**. Minimal competency represents professional expertise (in the case of voluntary certification programs) or a level of recognized proficiency enabling protection of the public and maintenance of the standards of practice (in the case of licensure/registration programs). The purpose of an examination program is **not** to identify and challenge the most knowledgeable candidates. While no profession likes to think of its recognized registrants as only *minimally competent*, understanding this concept is necessary to establish the criteria against which to measure aptitude. Workshop discussion will include identifying the attributes of a minimally competent practitioner: What skills are required to practice safely, to serve and protect the public? Where is the point of demarcation that distinguishes the competent from the incompetent practitioner?

Another critical concept that must be explored is that of **entry-level** practice. Based upon the eligibility requirements and knowledge, skills, and abilities associated with minimal competency,



we can establish a profile of the entry-level practitioner. That is not to say that the professional is new to the vocation; but he/she only JUST qualify as eligible and competent. Therefore, if one eligibility requirement states that candidates must have three years on-the-job experience, then our prototypical candidate should have just three years' experience, neither more nor less.

Once the profile of the candidates – entry-level practitioners, having just met the education and experience requirements – is established, the SMEs will examine the knowledge base that these candidates would be expected to have. It is *not* appropriate to expect candidates to demonstrate competence in areas to which they might be exposed only after significant experience on the job, or through voluntary continuing education or research, unless those criteria are required to sit for the examination. The candidate's **opportunity to learn** establishes the **relevance** of the knowledge base candidates would be expected to have.

Another activity focuses on the examination content outline representing the knowledge, skills and abilities demonstrated by the minimally competent, entry-level practitioner. SMEs will be asked to review which content areas might prove most challenging for candidates, and why. This concept of **difficulty** is distinct from the concepts of task importance and frequency of performance. It may be important for all entry-level candidates to perform a task (CPR for example) but the task (e.g. learning to perform CPR) may not be a particularly challenging or difficult one. Candidates may be asked to perform a task frequently, but that task may prove highly demanding. SMEs will be asked to keep in mind which content areas might prove most difficult or challenging. This assists SMEs with the understanding of the minimally qualified candidate within the content of the examination.

The Standard Setting Process – Research and Ideology

Passing standards are based upon either **criterion-referenced** or **norm-referenced** measurement models. **Criterion-referenced** measurement model compares candidate's performance against defined criteria of proficiency (standards). In other words, is a candidate able to perform specific delineated tasks and demonstrate a defined set of knowledge, skills and abilities? Establishment of a legally defensible standard (passing score) for certification and licensing credentialing relies upon criterion-referenced measurement; a standard which is fixed in difficulty across time and examination forms.

By contrast, interpretation of a candidate performance based upon comparison with a group of individuals (fellow candidates, for example) is called **norm-referencing**. "Grading on a curve" – passing or failing a set percentage of test takers based upon the distribution of scores – is an example of norm-referencing. For certification and licensure, setting a standard using a norm-referenced model is inappropriate; credentialing must be based upon individual candidate's performance measured against a set of criteria, separate from the abilities of a candidate group. For example, for a group of highly-able candidates, a norm-referenced model setting a 65% pass point would arbitrarily pass 65% percent of the candidates, despite the fact that more candidates in the group may be competent. Similarly, for a less-able candidate group, 65% would still pass with a norm-referenced model, even if none of the candidates were minimally competent.



By establishing criteria required of the minimally competent entry-level practitioner and judging candidate performance against those criteria, we have confidence that if all candidates who are tested are competent then **all** candidates will pass. Likewise, if all candidates tested fail to demonstrate competency none will pass.

There are numerous protocols for establishing a criterion-referenced standard, with a variation of the **Angoff Method** most widely used and recognized within the profession testing industry. This model is named for William H. Angoff, a pioneer in the field of psychometrics, and relies on SMEs using *empirical* (experience-based) *knowledge* to make individual evaluative judgments about how candidates will respond to individual test questions. The process further asks the SMEs to review these decisions as a group to evaluate and reach consensus on the decisions that are made.

Collecting Ratings – Setting the Standard

The first exercise asks the SMEs to draw back upon the earlier discussion of the entry-level, minimally competent candidate. For each question on the examination, the SMEs will be asked to envision a group of 100 entry-level minimally competent candidates and to predict (based upon empirical knowledge) what percentage of this group would answer the first item on the examination correctly. Alternatively, the SMEs can provide ratings based on the probability of a single minimally competent practitioner answering the question correctly. It has been our experience the first rating style is most effective.

A discussion of test taking probability statistics is integral to this exercise. SMEs are asked to keep in mind that random guessing on a multiple choice examination with four options will result in approximately 25% of candidates answering an item correctly. Even the most difficult items would not be expected to have a rating below 25%. Likewise, even the most capable candidates make clerical errors when taking tests: Even a very easy item is unlikely to have 100% of candidates choosing the correct response. Performance or practical examinations require the demonstration of a set of observable criteria, so the “guessing” factor is less applicable; candidates either fail or succeed in successfully performing a task. But, even the most able candidates are affected by outside factors, such as anxiety, interpretation of directions, time restraints or equipment variations, and it is unlikely that 100% of candidates would correctly perform an individual task on a performance examination. Therefore, we expect ratings to fall in a range of 25% to 95%.

Once the first rating is made, the values from each SME are entered into a spreadsheet for evaluation and discussion. The key is then shown to the SMEs and the SMEs are reminded that if they personally would have answered incorrectly, that they may want to assign a lower rating – indicating that few candidates would answer the item correctly. In addition, the p-value (proportion of candidates answering an item correctly) is shown to the SMEs to provide impact data. For example, if an item with a p-value history of 0.35, a panel decision to award an Angoff rating of 0.85 would be unsound. Next, disparity among judgments (indicated by values falling outside the mean and a higher **standard deviation**) is discussed. The object of the discussion



is to insure that all SMEs understand the Angoff process and that the panel is able to reach consensus. SMEs whose judgments consistently fall outside the mean distribution may be considered **outliers** and depending on post-meeting statistical analyses (i.e., rater consistency studies), their judgments may not be usable for standard setting. SMEs are afforded the opportunity to record comments about items, but are discouraged from using the exercise for in-depth review and discussion.

SMEs will be asked to make judgments on the next series of questions (e.g., next ten) on the exam. Again, judgment data will be entered, evaluated and discussed as a panel, until judgments have been collected for each item on the examination. Once judgments for all items have been collected, a final passing standard score is calculated. Discussion follows to ensure that this standard is appropriate, including review of previous standards and pass rates. Once consensus is reached a **raw score** representing the number of items a candidate must answer correctly is adopted.

Adopting a Scaled Score

Most testing programs today do not report **raw scores** as the passing standard; instead they adopt a **scaled score** for reporting purposes. Scaling is a means of “transforming” a score so that it is more easily interpreted and can be reported consistently across time and examination forms. Score scaling can be illustrated by thinking of the Fahrenheit and Celsius temperature scales. The temperature outside is the same regardless of whether it is reported in degrees of Fahrenheit or Celsius. Test performance is the same whether the score is reported as a raw score or a scaled score.

Legally defensible examination programs must have a means of differentiating between form difficulty and candidate ability. In other words, candidates may neither be *advantaged* nor *disadvantaged* by the ease or difficulty of an individual examination form. To insure that all candidates are treated equally, psychometric processes — such as equating and the use of item response theory — are used with Angoff cut score data to test and score examinations fairly and defensibly.

The standard setting process outlined in this document brief sets a standard for a single form of an examination; however, no matter how hard test developers try to create forms of identical difficulty, subsequent test forms will differ in difficulty, and that disparity must be taken into consideration when a criterion or passing standard is established. Because form difficulty is variable, passing points tend to shift slightly (up or down) for subsequent examinations. This means that the raw score passing point for one form might be 70, but for another form the raw score passing point may be 72. Reporting this raw score difference is often confusing for candidates; by adopting a fixed scaled score (for instance 70) and converting raw scores for individual forms to this scale, reporting is consistent and less confusing for candidates. The table below gives an example of a scaled score model for two forms of a 12 item examination.



Raw scores Form #A (used to set standard)	Scaled Score (adopted for reporting)	Raw Scores Form #B (subsequent form, easier than Form A)
2	30	5
3	40	6
4	50	7
5	60	8
6 (raw score to pass) ↘	70 Passing Standard	↙9 (raw score to pass)
7	80	10

The standard was set on Form A as a raw score of 6 and reported as a scaled score of 70 (or 70%). Form B was easier than Form A; the raw score reflecting competency was 9. Adopting a scaled score model allows reporting of the scaled score of 70 for both forms of the examination, creating a mechanism for addressing differences in form difficulty. Based upon the program design, SMEs may be asked to discuss the adoption of a standard scaled score to be used for reporting purposes.

Modification and Additional Tasks

For some programs the Standard Setting Workshop may include collecting Angoff judgments for individual test items that do not appear on the examination form used to set the standard. Having a bank of “spare” items with associated Angoff values affords programs testing small numbers of candidates (less than 50 a year) the ability to create additional examination forms with legally defensible passing standards. The judgment collection process is identical to that used to establish a standard for the exam, except that it is done on an item-by-item basis, rather than as a study of the collection of items making up one examination form.

Scheduled Review

Once a standard is established, its appropriateness must be periodically evaluated. Any time a Job Analysis Study reveals shifts or changes in the criteria required of the minimally competent practitioner, these shifts will alter examination content, examination items and examination forms. Depending on the profession, Job Analysis performance is likely to be recommended every 4-5 years. Based upon these changes, passing standards must be re-established as part of program support and maintenance. When it is not possible to update Job Analysis data, testing protocol recommends periodic reviews of established standards.

Setting the Standard: A Guide for Subject Matter Experts written by Schroeder Measurement Technologies, Inc., March 2005