Developing a Certification or Licensure Exam

By

Reed A. Castle, PhD
Director of Research and Development

Schroeder Measurement Technologies, Inc.
2494 Bayshore Blvd
Dunedin, FL 34685
http://www.SMTtest.com
727-738-8727

Fall, 2002

What is the goal of a credentialing examination?

The goal of a credentialing examination is to identify individuals who exhibit a predetermined level of competency. The examination scores are, in part, used to determine who is credential worthy and who is not. Test score validity is therefore of primary concern. In general, we want to evaluate a candidate's proficiency level against a standard that differentiates competence from incompetence. The inference made from a credentialing examination score is, "Did the candidate pass or fail the examination?"

In most instances, we are ***not*** concerned with how candidates perform relative to other candidates, but rather we are concerned with how a candidate performed relative to a passing standard. In other words, did a candidate exhibit a proficiency level at or above a minimum standard? Equally important is that examination scores should not be used for alternative purposes. For example, employers should not compare passing candidate scores for hiring or promotion.. Unless a validation study using exogenous variables (e.g., work performance ratings) has been conducted, and a validity argument has been made, comparing candidate scores is dangerous. Very rarely are these studies conducted, and when they are conducted, it often the case that the exogenous criterion variable does not exhibit reliability. To that end, Certification and Licensure Examinations, when properly developed, identify and classify candidates into one of two levels of proficiency resulting in a pass and fail decision based on their test score.

What are the steps in developing a defensible credentialing examination?

The remainder of this paper describes credentialing examination development topics in very general terms. It should be noted that there are variations to each of these steps, and for the sake of brevity, I will not discuss these variations. I have collapsed test development requirements into six basic steps.

1. Job Analysis/Role Delineation/Test Blueprint
2. Item Writing
3. Item Review
4. Test Development
5. Standard Setting
6. Statistical Analysis, Scoring, Equating & Scaling

Step 1. Job Analysis/Role Delineation/Test Blueprint

The foundation of a strong program is a job analysis (JA) study. It is the study that helps establish a link between test scores and competency for a particular profession. Specifically, the scores derived from a written examination should be content valid so that the pass and fail inferences are appropriate. Standards 14.10 and 14.14 in *The Joint Standards for Educational and Psychological Testing (AERA, APA, and NCME, 1999)* state the following:

> 14.10
> When evidence of validity on test content is presented, the rationale for defining and describing a specific job content domain in a particular way (e.g., tasks, knowledge, skills, abilities or other personal characteristics) should be stated clearly.
>
> *Comment: When evidence of validity based on test content is presented for a job or class of jobs, the evidence should include a description of the major characteristics that a test is meant to sample, including relative frequency, importance, or criticality of the elements*
>
> 14.14
> The content domain to be covered by a credentialing test should be defined clearly and justified in terms of importance of the content for the credential-worthy performance in an occupation or profession. A rationale should be provided to support a claim that the knowledge or skills being assessed are required for credential-worthy performance in an occupation and are consistent with the purpose for which the licensing or certification program was instituted.
>
> *Comment: Some form of job or practice analysis provided the primary basis for defining the content domain. If the same examination is used in licensure or certification of people employed in a variety of settings and specialties, a number of different job settings may need to be analyzed. Although the job analysis techniques may be similar to those used in employment testing, the emphasis for licensure is limited appropriately to knowledge and skills necessary for effective practice.*

To ensure that these and other standards are met, a comprehensive JA is required. The goal of the JA is to define the content that is valid for assessment. Further, a JA, along with appropriate test development procedures, helps ensure the legal defensibility of the resulting content and score inferences.

The first phase of the job analysis process includes conducting an exhaustive review of the role of the occupational professional. A comprehensive approach is used to develop an exhaustive list of tasks or behaviors to be included for evaluation by survey respondents. Documents that help are performance appraisals, job definitions, and relevant work-related literature and curriculum-related documents. Next, the committee of Subject Matter Experts (SMEs) evaluates which tasks are appropriate for inclusion on the survey. It is important to note that we are trying to identify behaviors associated with current practice. Therefore, we are not overly concerned with future professional issues as these are not currently measurable and we are not concerned with content that has become obsolete.

After the task list is generated, it is placed into a survey format and rating scales and demographic questions are added. Many rating scales can be used for a job analysis survey. I believe minimally that there are two ratings that one should consider for each

task: importance and frequency. In other words, there are two attributes against which each task should be measured. They are the importance of the task to the profession and how frequently the task is performed. The demographic questions on the survey serve two main purposes. First, the demographic questions help describe the respondent sample which helps evaluate the representativeness of the sample; and second the questions are used to further breakdown the data for evaluation.

After the survey is developed, it is distributed to a well-defined sample of practitioners. Efforts should be made to ensure that the return sample is reflective of the population of practitioners. Typical response rates can range from 20 to 40%. However, it is the number of responses that is more important that the percent of responses since the subsequent statistics will have less error as the sample size increases. Finally, data is tabulated and analyzed and decision criteria are used to determine which tasks are included in the test blueprint and how much weight should be assigned to various parts of the test blueprint.

Step 2. Item Writing

The second step in developing a defensible examination program occurs after the content (test blueprint) is delineated from the JA. After a test blueprint is developed, items are written to match it. This is the link between the profession and the examination. Each item is "linked" to a content area and each item on the examination must match content on the test blueprint.

A panel of SMEs is convened for this task. A psychometrician trains the item writers on how to write multiple-choice items or other item types. The goal of the training is to define "good" item formats, train the SMEs on what is not a good item (e.g., using none of the above as a distractor, making the key longer than the distractors, using negatively worded items), and review various item types and cognitive complexities associated with items.

As novice item writers start the process, it is important that the psychometrician review their work. This interactive review will help novice writers become more proficient. Good training and workshop coordination may increase the value of the item-writing workshop and prevent the squandering of resources in subsequent steps.

Step 3. Item Review

Once items are written and edited for format, a panel of SMEs will meet with a test developer to review the new items. This thorough process cultivates decent examination questions and subsequently good examination forms. The item review may consist of:

The Stem-

- ✓ **Does the stem present a clear and focused problem or stimulus to the candidate?**
- ✓ **Is the stem free of excess wording?**
- ✓ **Does the stem contain shorter statements instead of long and sometimes difficult to read statements?**
- ✓ **Is the stem free of blank spaces?**
- ✓ **Is most of the phrasing in the stem, in order to avoid redundant information in the options?**
- ✓ **Is the stem positively worded?**

The Options -

Distractors -

- ✓ **Do all options flow with the stem?**
- ✓ **Do any of the options contain "all of the above" or "none of the above"?**
- ✓ **Are the options phrased positively?**
- ✓ **Are all options free of terms such as "never" or "always"?**
- ✓ **Are all options plausible (do they relate to the stem conceptually)?**
- ✓ **Are all distractors real?**
- ✓ **Are all options free of humor?**
- ✓ **Are the options mutually exclusive/independent?**
- ✓ **Are all options of similar length?**
- ✓ **Are options ordered logically?**
- ✓ **Are distractors attractive to the uniformed candidate?**

Key-

- ✓ **Is the key the single best answer?**
- ✓ **Does the key erroneously contain information from the stem?**
- ✓ **For multiple items, is the key varied so it is not always the same letter?**

The Whole Item-

- ✓ **Is the key provided?**
- ✓ **Has the item been linked to the content outline?**
- ✓ **Has the item been correctly referenced?**
- ✓ **Is the item at the appropriate cognitive level?**
- ✓ **For situation sets, are the items independent of one another?**

As the test developer reviews items, it is important to activate items that meet the needs of the examination blueprint. For example, if section one of the test blueprint requires ten items, then activating 40 items in this content area is not as important as activating the required number of items in other content area sections.

Step 4. Develop a Test

Developing a test form (in our lexicon, pulling a form) is both an art and a science. The art component includes using items that do not cue one another (SME support is needed for this), placing scenarios in serial order, and ensuring that all graphics are correct and easily viewable to the candidates. The science component is similar to solving a puzzle since there are many constraints associated with pulling a valid examination. These constraints include:

1. Ensuring the items on the form match the test blueprint
2. Ensuring there is not great overlap with the previous examination form(s) (most programs specify the maximum percent overlap)
3. Ensuring items are selected to match the statistical requirements:
    a. Anchor test for equating
    b. Test form difficulty and score variability is similar to previous forms
    c. Information is maximized near the cut-score
    d. Items that perform poor statistically are not present
    e. Pretest items are imbedded in a rationale manner
    f. For IRT (Item Response Theory) Pre-equating, developing a test form around a specified cut-score

After a test form is created, it is extremely important that multiple SMEs review it. The SMEs are responsible for identifying items that cue one another and items that are highly similar in content. SMEs also review the items one more time as a safeguard against having poor items on the examination form.

Step 5. Conduct a Standard Setting Study

All new examination programs must conduct a study to determine a cut-score. While there are many methods used for establishing a cut-score, I am going to limit my discussion to a modified Angoff approach. Most testing programs currently use some form of the Angoff technique (Plake, 1998).

In general, a modified Angoff method employs expert judges who make an inference about minimally competent candidates' ability answer each item correctly on an examination. A panel of SMEs is used to help determine cut-score. After a detailed discussion of the borderline or minimally competent group,

SMEs are asked to provide an estimate of the proportion of minimally competent practitioners who would answer the item correctly. The process is repeated for each item across multiple rounds. The average rating (across judges) for each item is then summed to arrive at a passing score. While there is healthy debate within the measurement community about the veracity of the cut-score derived from a modified Angoff method, there are some practical virtues of this method. It is easy to incorporate, relatively less time is needed to implement, and it is computationally simple (Berk, 1986). Alternatively, SMEs may have difficulty grasping the concept of minimal competence.

As with other steps in the examination development process, standard setting is a fundamental component in establishing a claim of valid test score inferences. Great care should be taken during this step since this process defines the decision to pass or fail a candidate .

Administration

I did not include administration in my list because it is not a "test development" issue. However, it is a very important component of the testing process. It is the first time that a candidate interacts with the organization or one of its agents. To that end, it is very important to have well trained administration staff. Examination security and the prevention of cheating are other issues that should be of concern.

6. Statistical Analysis, Scoring, Equating & Scaling

Once an exam has been administered, it is important to conduct a test/item analysis. The purpose of this analysis is to identify any items that exhibit problematic statistics. In general, two indices are calculated. The first index is item difficulty and it is an expression of how easy or difficult the item was on that administration. Items that are extremely easy contribute little to measurement precision as do items that are extremely difficult. For this reason, these items are flagged for SME review. The second index is item discrimination. This index tells us the relative degree to which the item discriminates able candidates from non-able candidates. Items that exhibit low, no, or negative discrimination are flagged for SME review. In addition, most analyses provide additional information (e.g., meant test score by item option and proportion selecting a given option).

After final scoring is determined, a process called equating is conducted. Equating is the process of ensuring that a passing score is fair and consistent regardless of test form difficulty or candidate group ability. There are two general times which equating can occur. Pre-equating occurs prior to an administration, allowing for instant score reports as with Computer Based Testing. Post-Equating

occurs after final scoring. Both equating techniques require a psychometrician to conduct the study and to evaluate the results.

Finally, scaling is conducted. Scaling is the process of reporting scores from various test forms on the same scale. Because cut-scores may vary depending on the test form, it is important to report test scores on a common scale between forms. For example, the scaled cut-score for a given program may by 75, however form A may have a raw cut-score of 85 and form B may have a raw cut-score of 82. To help maintain a source of consistency, all raw scores are converted to a constant scale. Equating is the process that ensures the passing point is fair regardless of test form difficulty and candidate proficiency.

Summary

Developing a defensible examination program requires deliberate steps. These steps must monitored by someone with the expertise and experience to ensure that the program limits risk. This discussion was limited to traditional pencil and paper test development and administration model. Variations of computer-based testing will change some of the parameters and steps listed above and may be the topic of a future paper. Feel free to contact the author at rcastle@smttest.com with any test development or psychometric questions.

References

Berk, R. A. (1986) A consumer's guide to setting performance standards on criterion referenced tests. Review of Educational Research, 56, 137-172.

Plake, Barbara S. (1998) Setting performance standards for professional certification and licensure. Applied Measurement in Education, 11(1), 65-80.