Some Important Concepts in Test Development

The following information was provided by Stephen Johnson, PhD of rpm-data.
Dr. Johnson provides psychometric consultation for BCPE. BCPE follows the concepts described.

Certification

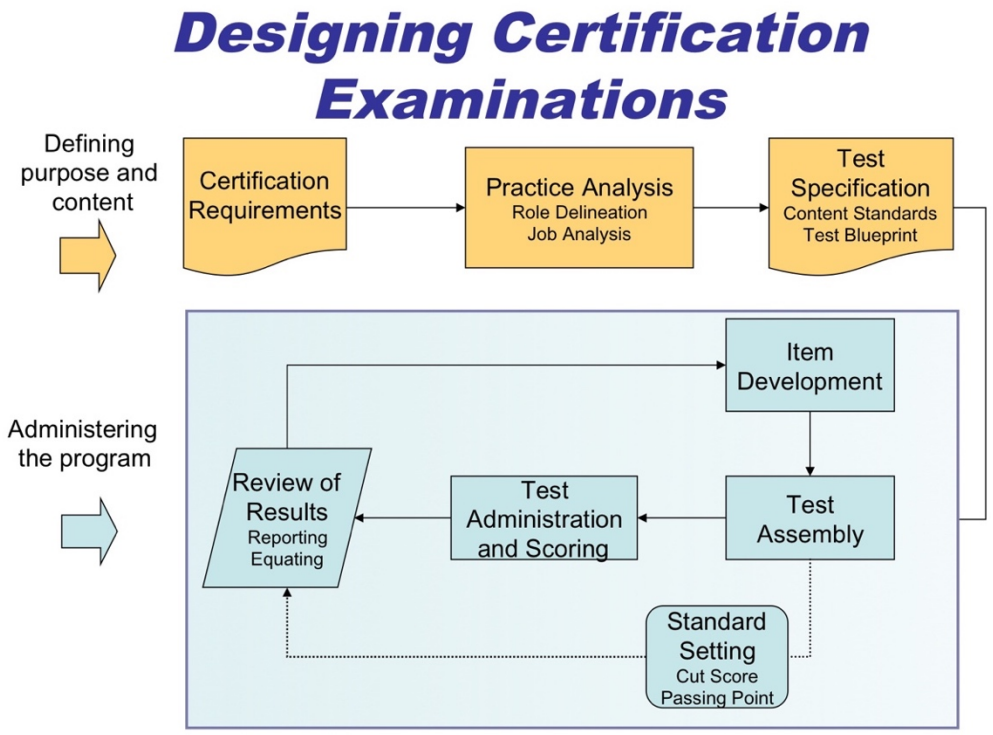
Certification refers to a process by which occupations not regulated by governments, identify whether a person has the required knowledge and skills necessary to perform the duties of that occupation. Certification is typically a process that includes assessment of educational qualifications, experience, and attainment of a passing score on one or more assessment tools.

The certifications are designed to distinguishing individuals capable of practicing safely and effectively from those who cannot. As a result, it is essential to ensure that certification examinations are valid, reliable, and objective measures of candidate ability.

Standardized Assessment

In order to develop a credible and valid examination, certification groups develop assessments following widely accepted standards and regulations (e.g., Standards for Educational and Psychological Testing, American Educational Research Association, 2014; Uniform Guidelines on Employee Selection Procedures, EEOC, 1978. They may also be accredited by a third party to the standards outlined under ISO/IEC 17-24:2012 which details the general requirements for personnel certification.

The diagram below describes in broad steps the workflow in the development of a standardized assessment.



The following pages describe in more detail a few of the important themes associated with this flowchart.

Role Delineation Study

A RDS, also known as a Job Task Analysis or Practice Analysis, is one of the methods used to identify and prioritize a clearly delineated set of domains, tasks, and associated knowledge and/or skills necessary to carry out the responsibilities of the job to the standards required for certification.

Most standards for the accreditation of certification programs (e.g., American National Standards Institute [ANSI], Institute for Credentialing Excellence [ICE]) require a demonstrable linkage between test specifications and the data collected during a RDS.

Information from the RDS is used at all stages of examination development, including the creation of test specifications, item writing, and examination construction. It is not possible within the constraints of most assessments to assess ALL the tasks, and their associated knowledge and skills, required in a field. RDS for certification examinations are concerned with a subset of a typical practitioner's job, that is, those aspects that are currently being performed, are critical to function at the level being tested, and can be assessed using the format of the examination. We generally refer to the identified tasks as *core competencies*.

A typical study consists of several critical elements; reviewing of existing material, discussions with a representative group of subject matter experts, and validation of their work.

A representative panel of subject matter experts (SMEs) well versed in the requirements needed to perform the job for the level of certification under consideration is empaneled. SMEs are typically certified, have relevant experience in the field, and represent the job or profession in criteria such as: workplace setting, job title, educational level, geographic region, ethnicity, and gender. Some SMEs may not work directly in the field but have experience working with potential certification candidates.

Core competencies identified by the panel must be:

1. Critical to the role of a recently certified competent professional – failure to perform them will result in harm to clients or other stakeholders;
2. Performed frequently – they are used in a variety of situations on a regular basis;
3. Somewhat unique to the roles and responsibilities of a competent professional (we need to focus on the knowledge and skills that cannot be assessed using other tools);
4. Can be tested in the format used for the examination; and
5. The majority of candidates would have been exposed to some level of formal or informal education/training necessary to learn the material.

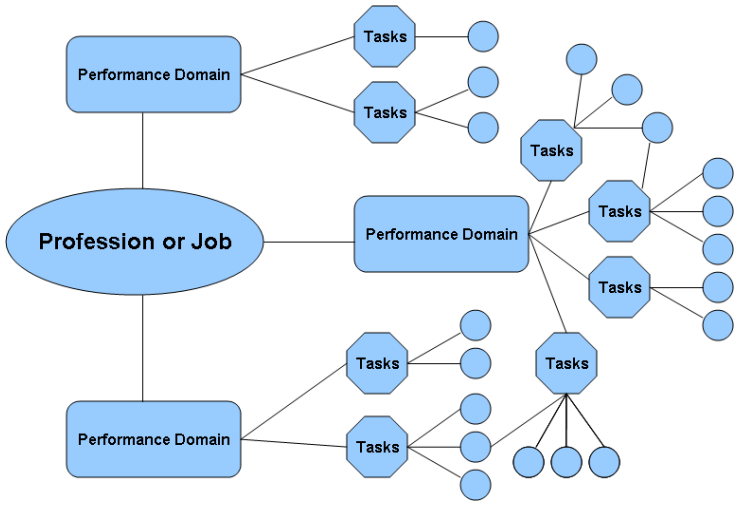
The panel's work is commonly validated by a survey of current certified practitioners and individuals providing services or performing the work consistent with the purpose of the credential. Survey respondents, like the SMEs, represent the job or profession. The respondents are asked about how often they perform the task, and whether it is critical to competent practice within the first three years after certification. The respondents are also asked to assess whether the role of a recently certified professional was adequately covered by the task list.

From this study a content outline is constructed. This defines what areas will be assessed, and ensures appropriate weighting in the scoring of presentation of content areas defined through the RDS process. Accreditation agencies require systematic re-evaluation of the content outline to ensure that it remains appropriate. As part of an organization's ongoing commitment to developing and maintaining a quality certification program, a role-delineation is typically conducted every five to seven years.

Defining Domains, Tasks, Knowledge and Skills

The structure used by the many fields consists of major areas of responsibility (domains), under which are critical and assessable tasks with their associated knowledge and skill statements.

Most content outlines specify four to eight domains. Tasks are the individual functions, whether mental or physical, required for certain aspects of a job/profession. Knowledge and/or skills (KS) are about how to do a task. They include information, actions, or other learnable and assessable constructs a candidate must possess in order to perform the tasks.



Notice that the domains and tasks are mutually exclusive. That is, a task cannot be associated with more than one domain, and domains are also independent. KSs, however, can be related to different tasks.

Item Writing

Certification agencies have many mechanisms to develop items. Many agencies hold item-writing workshops in which a panel of subject matter experts are brought together and given training and time to write items.

It is important to establish early in the development phase where an item fits into the test specification, what the correct response and incorrect options (distractors) are for the item, and the reference(s) that support the correct response for the item.

After an item has been written, it requires review from subject matter experts and an editorial review by a panel of subject matter experts and other appropriate persons. Panel members consider the:

- Appropriateness of the items for the certification;
- Match of the items with the framework and test specifications; and
- Quality of items.

After review by panel members, items are then revised and submitted for approval of inclusion on a test form by an examination development committee. Typically newly developed items are pre-tested before inclusion as scored items.

Pre-testing items

For established test programs there is an ongoing need to assess how newly written items work before they are placed as scored items on a test form. With the demand for faster scoring it is harder to meet the requirements for adequate statistical, professional, and fairness review of new items if the majority of a test form consists of new items.

A method used in many programs is to have each version of the test being administered to consist of scored items - items that have already been tested, reviewed, and approved for use as scored items – and pre-test or field-test items – items that are not scored but are being evaluated. This pre-testing typically involves anywhere from 10% to 25% of the items on a test form.

For the information about the pre-test items to be accurate, it is important that the items be tested on a representative group of the test-takers who are motivated to do as well as possible on the pre-test items. To achieve this, developers:

- Place pre-test items on operational test forms; and
- Do not reveal which items are scored and which are being pre-tested so test takers are motivated to try their best on every item.

Pre-test items that pass the fairness and statistical review with minimal or no changes are then available to be used when new test forms are developed. Items that have been substantially modified must undergo an additional round of pre-testing before they can be used as scored items.

Establishing performance standards

How do we know when someone has passed an examination? Examinations used for credentialing are held to the standard that they use a criterion-reference test. These types of tests compare a test taker's performance against an objective performance standard rather than other test takers. This requires that a test taker show that they know something rather than how they compare to other people. Upon successful completion of a criterion-referenced test, the test taker will have provided a demonstrable link between the purpose of the credential and an established standard of competence. This standard of competence, or performance standard, is typically defined as a cut score.

Beginning in the 1950s, methods were developed to establish the cut scores of an examination. A cut-score is the score required by a test taker to pass an examination. These techniques are typically referred to as standard setting. These methods define the minimally acceptable level of competence required by a test taker, and evaluate each question or an entire test in light of the level of competence required.

Within the certification field a standard setting panel of subject matter experts (SMEs) review what is known as the anchor version of a test, that is, the standard to which later versions will be anchored to. The SMEs are carefully chosen to reflect the profession, and also require intimate knowledge of the variety of test takers (e.g., through supervision or education of potential test takers). The SMEs make assessments of how difficult a minimally acceptable test taker would find the test.

There are many different ways to undertake a standard setting study but the most common approach is the Angoff family of techniques. For these methods the SMEs are trained to:

- Reach a common understanding of the purpose of the examination (including who is taking the exam);
- Reach a common understanding of the standard of minimum competence consistent with the purpose of the examination; and
- Estimate the difficulty of each question for the minimally acceptable test taker.

SMEs typically review the test two or three times and review the consequence of their decision. The purpose of this is to refine the accuracy of the identified passing standard. There is no desired percentage of successful test takers, nor is there a fixed percentage of correct answers that influence the standard.

Building tests with similar content and difficulty

Over time different test takers receive different sets of test items (called test forms). Test development agencies work hard to ensure that all test forms are comparable in content and difficulty to ensure that test takers who take different test forms on different dates are treated equally. This is achieved through a two-step process.

First, the different versions of an examination are built to a standardized test specification. Using a test specification ensures that all new versions of the test are comparable in content coverage, that test content is weighted in accordance with the requirements for competent performance in the profession, and that all forms of the examinations adhere to standards for content validity.

The test developers try to make the questions on different forms equally difficult, but more often than not, some forms of the test turn out to be harder than others. To adjust for these differences the tests are "equated." **Equating** is a family of statistical processes, but most programs use the Common Items Non-Equivalent groups design. This design compares the performance of one group of test takers on one test form to another group of test takers on another version. This process is used to compare the performance on later test forms with the performance of test takers who took the test form that was used to establish the passing standard.

To do this, the two sets of test takers are assessed for how well they performed on the items in common between the two test forms being compared. The assumption is that the two groups should have similar performance on the common items, and that any difference in difficulty between the test forms can be accounted for by the difficulty of the items unique to the new form. Adjustments can then be made to account for any difference in overall performance.

Scaled scores

Since test forms are possibly of different difficulty, providing raw scores can be misleading. As a result many programs use scaled scores, including the ACT® and SAT® examinations. Scaled scores are particularly useful at providing the basis for long-term, meaningful comparisons of results across different administrations of an examination.

Scaled scores are used because over the life of every testing program there are situations when:

- Changes in test length occur, for example, when a decision is made to assess more or fewer areas; but most often when the number of items that are scored versus pre-test (experimental) changes; or
- Different test forms of different difficulty are being compared.

For scaled scores, the passing standard (number correct) on any form of the examination is always reported as the same scaled score.

If we did not use scaled scores, people making assessment of scores on an exam would end up comparing apples to pears (or Celsius to Fahrenheit). Scale scores help test takers make easy comparisons by ensuring that all the scores are on the same scale (everything is in apples or Celsius).