1. Establish the test purpose

A good first step in the development of an exam program is to establish the test purpose. How will the test scores be used? Is the exam intended for certification or for licensure? Will the emphasis of the test be on minimum competency or on mastery of course content? Overall, will the test be low-stakes, moderate-stakes, or high-stakes for the examinees? That is, will the results of the test have important consequences for the examinees? The answers to these questions will have implications for many aspects of the exam, such as the overall length of the test, the average difficulty of the items, the conditions under which the test will be administered, and the type of information to be provided on the score reports. If you take the time at the beginning to establish a clear, focused test purpose, your goals and priorities will be more effectively met.

2. Conduct the job analysis

A job analysis is conducted in order to identify the knowledge, skills, and abilities that a professional in a particular field ought to have. Within a test development effort, a well-conducted job analysis helps provide for the validity of the test that is later developed. The job analysis contributes to test validity by ensuring that the critical aspects of the job become the domain of content that the test measures. Other, highly similar activities that are sometimes used to help promote the validity of a test include task analyses, practice analyses, and role delineation studies. A typical job analysis occurs in two phases. In the first phase a list of job-related tasks is generated and refined, while in the second phase a validation of that list of tasks is conducted.

3. Create the test specifications

After the overall content of the test has been established through a job analysis, the next step in developing a test is to create the detailed test specifications. Test specifications usually include a test description component and a test blueprint component. The test description specifies aspects of the planned test such as the test purpose, the target examinee population, the overall test length, and more. The test blueprint, sometimes also called the table of specifications, provides a listing of the major content areas and cognitive levels intended to be included on each test form. It also includes the number of items each test form should include within each of these content and cognitive areas.

4. Develop the initial pool of items

Once the test specifications are complete, the item writing phase of the test development project can begin. Typically, a panel of subject matter experts is assembled to write a set of test items. The panel is assigned to write items according to the content areas and cognitive levels specified in the test blueprint. Items are written for each of the item types identified in the test specifications. By far the most commonly used item type in standardized assessment is the multiple choice item, due to its relative advantages, including its ability to be used to measure at higher cognitive levels. Some exam programs use item specifications to further guide item writers with detailed requirements for each included item type. The total number of items that a particular exam program needs depends on specific aspects of the exam program. After the items have been written, they are stored electronically in an item banking software application.

5. Review the items

After a set of items has been written, an important next step in the test development process is to review the items. The items need to be reviewed for several different types of potential problems; thus it is often helpful to have different types of experts conduct specific reviews. Subject matter experts should review the items to confirm that they are accurate, clearly stated, and correctly keyed. Professional editors can then review the items for grammar, punctuation, and spelling. Measurement experts can review the items to be sure that they are not technically flawed. And the items can also be reviewed for fairness, to ensure that they will not be likely to disadvantage any examinee subgroups. The items may also be reviewed to ensure that they match the test specifications and are written at an appropriate readability level. The review process is valuable for identifying problems which should then be corrected before the items are field tested.

6. Field test the items

After a set of drafted items have been reviewed by test development and subject matter staff, the items should be ready for field testing on examinees. When items are field tested, they are not scored and they are not used to measure examinees; instead, the items themselves are evaluated. An item field test is conducted to collect information about how well the items are likely to function operationally, before they are actually used to contribute to an examinee's score. For that information to be accurate, it is important that the field test includes appropriate examinees. The field test examinees should be representative of the test population; that is, they should be as similar to future test-takers as possible. The examinees should also be motivated; that is, they should be attempting to do as well as possible when they respond to the items. Items may be evaluated at different test development phases, under pilot test, field test, and pretest conditions. In all these instances, data are collected to ensure the high quality of the items and tests. That information is then used to conduct an item analysis and review of the items, prior to allowing items to be included on operational test forms.

7. Assemble the test forms

After draft items have been reviewed and field tested, and item analysis statistics have been obtained on the items, operational test forms can be assembled. Each new test form for the exam program ought to be assembled to satisfy all of the elements in the test specifications, and particularly in the test blueprint. Items are selected from the bank to satisfy the content and cognitive proportions specified in the test blueprint, as well as to meet certain statistical criteria. This process helps ensure that each resulting test form satisfies the intended purposes of the test. Most exam programs follow this process to develop not just a single test form, but multiple forms. These multiple test forms should be assembled so that they are as similar to one another as possible, or parallel. The use of multiple, parallel test forms provides improved test security for the exam program. After test administration, statistical equating of the parallel forms is often needed.

8. Conduct the standard setting

Standard setting is the process used to select a passing score for an exam. Of all the steps in the test development process, the standard setting phase may be the one most like art, rather than science; while statistical methods are often used, the process is also greatly impacted by judgment and policy. The passing score (also known as the passing point, the cutoff score, or the cut-score) is used to classify examinees as either masters or non-masters. An examinee's score must be equal to or greater than the passing score in order for that examinee to be classified as a master, or to pass the test. If an examinee is misclassified, that is referred to as a classification error. Typically, the judges set the passing score at a score point on the exam that reflects the minimum level of competency to protect the public from harm or to provide minimal competency at the occupational level being assessed. For the standard setting to be conducted successfully, the panel of judges should be carefully selected and then thoroughly prepared and trained for their task. There are a number of approaches to standard setting including informed judgment, conjectural, and contrasting groups methods.

9. Conduct the item analysis

The item analysis is an important phase in the development of an exam program. In this phase statistical methods are used to identify any test items that are not working well. If an item is too easy, too difficult, failing to show a difference between skilled and unskilled examinees, or even scored incorrectly, an item analysis will reveal it. The two most common statistics reported in an item analysis are the item difficulty, which is a measure of the proportion of examinees who responded to an item correctly, and the item discrimination, which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area and those who are not. An additional analysis that is often reported is the distractor analysis. The distractor analysis provides a measure of how well each of the incorrect options contributes to the quality of a multiple choice item. Once the item analysis information is available, an item review is often conducted.

10. Administer the test

Test administration procedures are developed for an exam program in order to help reduce measurement error and to increase the likelihood of fair, valid, and reliable assessment. Specifically, appropriate standardized procedures improve measurement by increasing consistency and test security. Consistent, standardized administration of the exam allows you to make direct comparisons between examinees' scores, despite the fact that the examinees may have taken their tests on different dates, at different sites, and with different proctors. Furthermore, administration procedures that protect the security of the test help to maintain the meaning and integrity of the score scale for all examinees.

11. Score the test

Once a test has been administered and the results collected, examinees' responses can be scored. Examinee answer sheets are first scanned and then an item analysis is conducted. If the item analysis reveals any problems, the problematic items may be re-keyed or dropped from operational use. If these corrections are needed, the answer sheets are then rescored. Once scoring is complete, the passing score that was set for the exam is then applied. Next, it may be useful to compute the proportion of examinees who were classified as masters and non-masters, or to plot a frequency distribution of all examinees' total test scores. Finally, score reports for the individual examinees are prepared. Each examinee’s raw score (that is, the number or proportion of items that the examinee responded to correctly) is often converted to one or more other types of scores. Various types of scores that an exam program might choose to provide on the examinees’ score reports include the pass/fail decision, diagnostic information, the examinee’s percentile rank, and the examinee’s scale score.