

ITEM DEVELOPMENT: WHERE DO GOOD TEST QUESTIONS COME FROM?

Think of a test question as a product, like a disposable razor, but a lot more expensive and with more long-term impact on your life. The razor, made of plastic and metal, was molded, cut, sharpened, assembled, inspected, packaged, distributed, and finally, bought by you. You probably never thought much about the process that brought that razor to you, but that's because it's not all that important.

But a test question, that's different. It's important to people, at least at the moment they are trying to answer it correctly. And they are probably trying to understand it and evaluate it from the moment they see it until they are on to the next one.

Does a test question go through a similar development process that a razor does, from raw material to useful product? How was the question originally written? Or better yet, why was it written? What reviews and changes did it go through? How many people actually read it and agreed that it should be on the test? These are great questions (no pun intended) and deserve to be answered.

By the time a test question, called an item, is used in an actual test it has been authored, reviewed, edited, rewritten several times, edited again, field tested, and subjected to statistical scrutiny. Only the best items make the cut. This process of refining and culling usually takes several weeks and involves dozens to hundreds of experts (subject matter experts, editors, quality assurance testing specialists, psychometricians, beta test takers, etc.). With such intensive (and expensive) human attention, it's understandable when testing programs go to great lengths to protect their items with security procedures and reduce their unnecessary exposure with new technologies such as computerized adaptive testing.

First of all, a question should not be written until the knowledge, skill, or ability has been identified which the question should measure.

A test, for example, may be used to measure job knowledge in applicants for a job. Relevant knowledge might be: The candidate must have knowledge of tools to repair and change a tire.

(Job skills like this one are best identified by interviewing experts through a process known as a job task analysis.) Once the skill is identified, one or more test questions can be written with the goal of measuring the skill as well as possible.

An expert in the subject matter (SME) who probably has some experience in writing test questions is the first person to actually produce the first draft of the question, which may include graphics as well. As the question is being authored, often the SME will get help from colleagues to make sure the question is accurate and relevant. All questions, but especially multiple choice questions, require that the SME follow specific format rules for such questions.

After the initial authoring, the question, along with all the others produced, is sent to an editor. The editor is not an SME, but does understand the rules of language, style and the proper formatting of questions. The editor will fix the language and design problems with the sole goal of reducing ambiguity. For example, if the editor notices that, because of vague wording, two choices of a multiple choice question are correct (when only one should be), he or she will rework one of them or alert the original SME to the problem. The result is a better question.

The question is returned to a group of SMEs who review each one for technical accuracy, representation and relevance. One name for this process is the technical review. Does the question really measure the test objective? Is it an important question, measuring important knowledge? Does the test “need” the question to be balanced across the content domain? Is the question accurate, including a correct answer? The question is usually changed (and may even be deleted) at this stage.

Since changes were made during the technical review, the question is returned to the editor. As before, the editor will fix any obvious errors introduced during the technical review.

When all questions have been refined in this way, they are subjected to an actual “field test” of their quality. In what is called a beta test, questions are answered by actual candidates in circumstances that mimic the motivation and environment of a real test. Beta test participants can comment on the quality of each question, but their answers also create test results that are subjected to a statistical analysis. The analysis will catch those questions that aren’t performing properly, even when those questions have passed all quality checks to this point. These poorly

performing questions cannot be part of the final version of the test. Obviously the questions you see on the actual certification test survived this beta process.

The final set of questions then form the official test or exam. But before anyone can take the test, it goes through a final series of quality assurance steps. While these steps are focused on the actually functioning of the test, the questions are briefly reviewed once more.

These several steps make sure that each test question, while not perfect, is as good as it can be at measuring the identified skills, knowledge or abilities. With enough of these great questions it is possible to produce a reliable and valuable test score that indicates if a person has the necessary skills, knowledge or abilities to do the job or to receive a diploma.

To illustrate this procedure more specifically, let me lastly describe a possible way to thoroughly develop items for large scale testing usage.

Authoring. Subject matter experts (SMEs) with no item writing experience are usually invited to “workshops” where in a 4- or 5-day period they learn how to write good items and then, with mentoring help, produce enough items to create a test. By the end of the week, six to 10 SMEs will have authored a pool of 250 to 400 items. The pool of items, also called a bank, is then sent to editors.

Editing. Specially trained editors simultaneously look for 4 general kinds of problems with the items and then fix what they can. The general areas are:

- ✓ *Language.* For the language edit, the editor identifies and fixes problems in grammar, spelling, vocabulary and punctuation.
- ✓ *Psychometric.* The psychometric review will focus on item format, making sure that the various specific rules for writing items—and there are hundreds of them--have been followed. For example, does a multiple choice item have enough answer options? Or, is a graphic clear and positioned properly?
- ✓ *Gender Bias.* This review looks for components of an item (text or graphics) that would make it more difficult for one gender to answer the question correctly. Such problems may occur from time to time.

- ✓ *Cultural Bias*. Many exams are authored in English by native English speakers. It, for instance, is not uncommon for North American cultural references to be authored inadvertently. For example, the item might refer to a “government” agency that sets standards for Internet use. If the question is unclear which government is being addressed, test takers outside of the United States might assume their own government. Or they might assume that the question is referring to the US government, but would know very little about that government. Either way, this question is inherently unfair.

Accuracy Review. The items are then reviewed by several other SMEs who make sure that each item contains accurate and relevant content and reflects the performance objective being measured. The reviewers also confirm the correct answer. Items that ultimately do not have consensus approval of these SMEs are rejected. After another round of editing the remaining items are compiled into a beta test.

Beta Test and Statistical Analysis. A (computerized) test is created and offered to the testing audience as a field test of the items. Over a period of a few weeks, fifty to 200 participants are recruited to answer each question in the pool. When the beta period has ended, the testing results are collected and analyzed statistically. To be eligible for the final test, an item must perform within specific statistical limits. For example, it cannot be too easy or too difficult, and it must discriminate between competent and less competent candidates. Mostly, about one-fourth of the items do not survive the beta process and is abandoned.

From these high-quality surviving items final tests are produced. These items greatly enhance the reliability and validity of the test, allowing it to be an effective tool in the decision-making process on individuals.