Constructing Objective Measure of Learning: Tests, Exams and Quizzes

Introduction

Once faculty have identified that there is a need to use an objective test to assess students' learning and achievement in a class, there are several steps for test construction that will help ensure that the resulting measure is of high quality. These steps are listed below with a description of what the faculty should accomplish in each step.

1. Test conceptualization

In this step, the faculty determines if there is a need for an objective test or if some other method of formative or summative assessment will be better suited to the faculty's purpose. To help determine this need, the faculty must identify the content, abilities and thinking skills the test is meant to measure, as well as the purpose of the test. Will it be to formatively provide information about what instruction or practice is still needed or will it be used summatively to measure student achievement for grading purposes? Once the need and purposes for the test are determined, the faculty must specify the content and create the test.

2. Test construction

During the test construction process, the faculty will specifically define the content, abilities and thinking skills the test is meant to measure, determine the length and format of the test, as well as the format for students' responses (i.e., open-ended versus closed-ended items) and the scoring scheme for the test.

a) **Test content specifications table (Test Blueprint)**
   The test blueprint helps the faculty ensure that the course outcomes, instruction and test content are all aligned appropriately. The blueprint explicitly identifies the course outcomes covered within a test, the topic areas that will be addressed within each outcome and the type of cognitive response the test is meant to elicit. Will the items be written to simply have students recall or recognize content or will they require students to apply content to real-life situations? Will students be asked to analyze or make an evaluation of information based on concepts learned in the course? Faculty can use Bloom's Taxonomy of Educational Objectives to help determine the cognitive process desired for each item of the test.

Remember that it is easiest to write objective items at the lower levels of Bloom's taxonomy and more difficult to write those at higher levels. For this reason, the test blueprint is very useful in helping faculty stay focused on the cognitive skills a test is meant to measure when writing questions. Writing good questions is a time consuming process, so it is helpful to start this process well in advance of the testing date to avoid the pitfall of writing questions that only assess lower cognitive processes and may have other types of errors, such as grammatical or typographical errors or more than one correct answer.

If it is important to assess higher levels of cognitive processes that are too difficult to capture in objective test questions, such as items written at the evaluation and analysis levels. These cognitive skills may be better assessed with a different method, such as an essay or research paper.

An example of a test blueprint is shown below. Note that not all of Bloom's levels are being assessed with this quiz. The numbers shown in the columns indicate how many items will be included on the quiz for each topic area at each level. When considering the number of items and total test length, it is important to consider the length of the examination period. Students should have ample time to complete the test within the time allotted without feeling undue time pressure.

Quiz 1 Content: Text Modules 1, 2, 5, 6, and 16 and Schmitt (1996). Twenty items each worth 1 point.

Course Learning Outcome	Content	Cognitive Level		
		Knowledge	Application	Evaluation
Articulate the importance of psychometric analyses for demonstrating the usefulness of educational and psychological tests (7 items)	Measures of Central Tendency	1		
	Measures of Dispersion	1		
	Measures of Distribution	1		
	Definition of Reliability	1		
	Classical Test Theory/Latent Traits		1	1
	Sources of Test Error	1		
Interpret estimates of test reliability and validity and evaluate the usefulness of tests based on these estimates (4 items)	Reliability Estimate Interpretation	1	1	
	Judgments of Test Usefulness		1	1
Compare and contrast, as well as calculate, various estimates of reliability and validity (6 items)	Correlation Coefficients	1	1	
	Types of Reliability Estimates	2	1	1
Articulate the conditions under which each form of reliability and validity evidence is appropriately used (3 items)	Uses of Various Reliability Estimates		1	1
	Limitations of Psychometric Theory			1

b) **Format of the test**

Once the specific content, including topic areas and corresponding cognitive levels, is identified, the faculty must decide on the format on the test. The most frequently used closed-ended format is multiple choice, but other formats include matching, true/false and fill in the blank when students are given a list of options. Tests can also include constructed response items such as fill in the blank (when no word list is given), computation items, short answer items, word problems and essays.

With constructed response items, it is helpful to prepare a key with ideal answers ahead of time to ensure that students are able to respond to these items based on information

they should have learned in the course to date. These ideal answers and even rubrics can aid in evaluating the correctness of these items.

Directions should be included at the beginning of the test to indicate to students how to record their answers and how the test will be scored, such as "This test is worth 100 points, and each item answered correctly will be worth 2 points."

If the test includes several different formats, all items of a single type should appear together on the test with their own instructions. Students' scores may not accurately reflect what they know and can do if they have to spend time and energy constantly switching back and forth between item types.

Some guidelines for writing multiple choice items include the following:
- The question stem should be clearly written and free of typographical and grammatical errors and should reflect one concept. Double-barreled items are those that are trying to ask about more than one concept in a single question and should be avoided.
- Each distractor should be a plausible, yet incorrect option. When the distractors are too obviously not the correct answer, they are not distractors and can be easily ruled out.
- All answer options should be approximately the same length and have the same grammatical form to be consistent with the stem. For example, if one distractor is written in the past tense, all other options should also be past tense.
- The correct answer should not always correspond to the same letter choice. Test wise students will quickly pick up on this, and if they don't know the answer, they will always guess the same letter.
- "All of the above" and "none of the above" options tend to be throw away options, usually making it too easy for students to guess correctly. These should be used with caution.
- Avoid overwhelming students with additional options that combine two correct answers, such as "both b and c." If the intended goal of these questions is to make students correctly choose multiple correct answers from a list of possible choices or to select those that do not fit, a check-all-that-apply question is better than a single-option question. Many test scoring packages are now quite flexible in allowing questions that have more than one correct answer.

c) **Scoring the test**
It is important in advance to specify the scoring scheme of the test. How much will the test be worth overall? How much will each item be worth? This information should be communicated clearly on the test itself so that students know in advance how the test has the potential to impact their overall course grade and how much time to devote to

different sections of the test. It will also ensure that there are no unexpected surprises when scoring the tests or when students receive their results.

3. Test tryout

This step is usually skipped in the test construction process. However, it is important to have at least one other person attempt to complete the test before it is administered. This review will help to identify confusing directions or question text, minimize duplicative response options and questions that have no correct answer and ensure that the test can be completed in the allotted time.

4. Test administration

When the test is administered, the faculty should try to limit any factors that will prevent students from doing their best. The room should be quiet and at a comfortable temperature. All students should be advised to have what they need in front of them before the test begins. The faculty should try to adopt a calm and pleasant demeanor.

The goal during the administration is to maintain an environment where students are able to capably demonstrate what they know and can do. When students are feeling pressured to finish a test within the allotted time, when the test format is jumbled or confusing, when items are not clearly written or have many errors and when the room is too hot, too cold or too noisy, these factors will impact test scores in a negative way so that students' scores are more reflective of their tolerance of imperfect conditions rather than their level of achievement of course learning outcomes.

In addition, if the faculty comes into the testing situation in a bad mood, seems anxious or expresses some other negative emotion, students often see this as reflecting that the test is going to be difficult or that they are not prepared. In either case, they are likely to do more poorly on the test than if the faculty is calm and pleasant.

Any questions or concerns raised by students about the test during the administration should be noted and reviewed after the tests are scored.

5. Test review

After the test is administered, item analysis should be performed to determine item difficulty and item discrimination for each item. Most scoring packages will produce item analyses if requested. For a multiple choice test, general guidelines suggest that item difficulty for each item should range from 0.50 to 0.80. Items with lower levels are too difficult and those with higher levels are too easy. (Although having a few easy items right at the beginning of the test is a good way to relieve students' initial anxiety so that they can perform to the best of their ability on the remaining items.)

The correct response on a multiple choice question should have an item discrimination value of at least 0.30, meaning that more students who overall did well on the test get the item correct. Each distractor should have a negative discrimination, indicating that students who overall weren't as prepared are selecting incorrect options.

Item analysis information and any issues raised by students during the administration can be used to revise the items so the items are more effective in the future. This item analysis information can also be used with the current test takers to help determine which concepts are still confusing and where the students might need additional instruction or practice.

Summary

The most effective tests go through a very rigorous development process to ensure that they are accurately measuring what students know and can do. Faculty need to be as planful about test construction as they are about every other aspect of instruction.

University at Buffalo
Center for Educational Innovation

5 Norton Hall, Buffalo, NY 14260

PHONE 716-645-7700

EMAIL ubcei@buffalo.edu

WEB buffalo.edu/ubcei