Handbook on Test Development:

Helpful Tips for Creating Reliable and Valid Classroom Tests

Allan S. Cohen

and

James A. Wollack

Testing & Evaluation Services
University of Wisconsin-Madison

1. Terminology Used in Testing

The following are some common terms used in testing:

Test Blueprint. The test blueprint (or test specifications) identifies the objectives and skills which are to be tested and the relative weight on the test given to each. This statement necessarily precedes any development of the test. These specifications provide a "blueprint" for test construction. In absence of such a blueprint, test development can potentially proceed with little clear direction. The development of such a set of specifications is the crucial first step in the test development process.

One must be mindful that the test specifications cannot and should not remain static. Pedagogy is not static and the specifications for each test need to be continually reviewed and modified to reflect the current state of knowledge.

Item Development. The term item is used as a shorthand for questions on the test. Item development can proceed only when a clearly agreed upon set of objectives is available. To as large an extent as possible, an item should measure only a single objective. Each objective, however, should be measured by one or several items, depending on the test specifications.

Item format. The format of the item necessarily proceeds from the test blueprint. The blueprint indicates the kinds of skills and the balance of test content to be measured. The selection of item types and test format should be based on the kinds of skills to be measured and not on some personal like or dislike for a particular item format. The use of multiple-choice questions, for example, may make sense for large group testing on knowledge of the mechanics of English. This type of item is not generally appropriate, though, as a direct measure of writing skill. If the intent is to determine whether an examinee can write a clear coherent essay, then an essay or free-response format is clearly more appropriate than a multiple-choice format. There is no inherent goodness or badness in any type of question format. The choice must be made on the basis of the behavior to be tested.

One issue which sometimes constrains the selection of test item format is the need for fast, relatively inexpensive scoring. In general, scoring fixed-response items, such as multiple-choice items, can be done faster and less expensively than scoring free-response items such as fill-in-the-blanks, short answer or essay items. This is particularly true when there are a large number of examinees whose examinations need to be scored quickly.

Many classroom objectives can be measured adequately with items that are amenable to machine scoring. There are also a number of objectives, however, which are more appropriately measured under other types of formats. Instructors are encouraged to use select the type or types of item formats which are best suited for measuring the desired skills.

1.1 Terminology Regarding Multiple-Choice Test Questions

Multiple-Choice Item: This is the most common objective-type item. The multiple-choice item is a test question which has a number of alternative choices from which the examinee is to select the correct answer. It is generally recommended that one use 4 or 5 choices per question, whenever possible. Using fewer alternatives often results in items with inferior characteristics. The item choices are typically identified on the test copy by the letters A through E.

Stem: This is the part of the item in which the problem is stated for the examinee. It can be a question, a set of directions or a statement with an embedded blank.

Options/Alternatives: These are the choices given for the item.

Key: This is the correct choice for the item.

Distractors: These are the incorrect choices for the item.

2 Guidelines for Developing Test Items

The following are some guidelines that you should use for preparing test items.

2.1 Writing Multiple-Choice Test Items

The general rules used for writing multiple-choice items are described below. Recognize that these are general rules; not all rules will be applicable to all types of testing.

1. The stem should contain the problem and any qualifications. The entire stem must always precede the alternatives.

2. Each item should be as short and verbally uncomplicated as possible. Give as much context as is necessary to answer the question, but do not include superfluous information. Be careful not to make understanding the purpose of the item a test of reading ability.

3. Avoid negatively stated items. If you have to use this kind of item, emphasize the fact by underlining the negative part, putting it in capital letters or using italics. (For test construction purposes, if possible, put all such items together in a single section and indicate this with separate directions.)

4. Keep each item independent from other items. Don't give the answer away to another item. If items require computation, avoid items that are dependent on one another.

5. If one or more alternatives are partially correct, ask for the "best" answer.

6. Try to test a different point in each question. If creating item clones (i.e., items designed to measure the exact same aspect of the objective), be certain to sufficiently change the context, vocabulary, and order of alternatives, so that students cannot recognize the two items as clones.

7. If an omission occurs in the stem, it should appear near the end of the stem and not at the beginning.

8. Use a logical sequence for alternatives (e.g., temporal sequence, length of the choice). If two alternatives are very similar (cognitively or visually), they should be placed next to one another to allow students to compare them more easily.

9. Make all incorrect alternatives (i.e., distractors) plausible and attractive. It is often useful to use popular misconceptions and frequent mistakes as distractors. In the foreign languages, item distractors should include only correct forms and vocabulary that actually exists in the language.

10. All alternatives should be homogeneous in content, form and grammatical structure.

11. Use only correct grammar in the stem and alternatives.

12. Make all alternatives grammatically consistent with the stem.

13. The length, explicitness and technical information in each alternatives should be parallel so as not to give away the correct answer.

14. Use 4or 5 alternatives in each item.

15. Avoid repeating words between the stem and key. It can be done, however, to make distractors more attractive.

16. Avoid wording directly from a reading passage or use of stereotyped phrasing in the key.

17. Alternatives should not overlap in meaning or be synonymous with one another.

18. Avoid terms such as "always" or "never," as they generally signal incorrect choices.

19. To test understanding of a term or concept, present the term in the stem followed by definitions or descriptions in the alternatives.

20. Avoid items based on personal opinions unless the opinion is qualified by evidence or a reference to the source of the opinion (e.g., According to the author of this passage, . . .).

21. Do not use "none of the above" as a last option when the correct answer is simply the best answer among the choices offered.

22. Try to avoid "all of the above" as a last option. If an examinee can eliminate any of the other choices, this choice can be automatically eliminated as well.

2.2 Writing Essay Test Items

Essay items are useful when examinees have to show how they arrived at an answer. A test of writing ability is a good example of the kind of test that should be given in an essay response format. This type of item, however, is difficult to score reliably and can require a significant amount of time to be graded. Grading is often affected by the verbal fluency in the answer, handwriting, presence or lack of spelling errors, grammar used and the subjective judgements of the grader. Training of graders can require a substantial amount of time and needs to be repeated at frequent intervals throughout the grading.

The following rules may be useful in developing and grading essay questions:

1. The shorter the answer required for a given essay item, generally the better. More objectives can be tested in the same period of time, and factors such as verbal fluency, spelling, etc., have less of an opportunity to influence the grader. Help the examinees focus their answers by giving them a starting sentence for their essay.

2. Make sure questions are sharply focused on a single issue. Do not give either the examinee or the grader too much freedom in determining what the answer should be.

2.3 Guidelines for Writing All Types of Items

Some additional guidelines to consider when writing items are described below:

1. Avoid humorous items. Classroom testing is very important and humorous items may cause students to either not take the exam seriously or become confused or anxious.

2. Items should measure only the construct of interest, not one s knowledge of the item context.

3. Write items to measure what students know, not what they do not know.

3 Guidelines for Review of Test Items

The following guidelines are recommended for reviewing individual test items. When you review an item, write your comments on a copy of the item indicating your suggested changes. If you believe an item is not worth retaining, suggest it be deleted.

1. Consider the **item as a whole** and whether
   a. it measures knowledge or a skill component which is worthwhile and appropriate for the examinees who will be tested;
   b. there is a markedly better way to test what this item tests;
   c. it is of the appropriate level of difficulty for the examinees who will be tested.

2. Consider the **stem** and whether it
   a. presents a clearly defined problem or task to the examinee;
   b. contains unnecessary infomation;
   c. could be worded more simply, clearly or concisely.

3. Consider the **alternatives** and whether
   a. they are parallel in structure;
   b. they fit logically and grammatically with the stem;
   c. they could be worded more simply, clearly or concisely;
   d. any are so inclusive that they logically eliminate another more restricted option from being a possible answer.

4. Consider the **key** and whether it
   a. is the best answer among the set of options for the item;
   b. actually answers the question posed in the stem;
   c. is too obvious relative to the other alternatives (i.e., should be shortened, lengthened, given greater numbers of details, made less concrete).

5. Consider the **distractors** and whether
   a. there is any way you could justify one or more as an acceptable correct answer;
   b. they are plausible enough to be attractive to examinees who are misinformed or ill-prepared;
   c. any one calls attention to the key (e.g., no distractor should merely state the reverse of the key or resemble the key very closely unless another pair of choices is similarly parallel or involves opposites).

4 Assembling a Test Form

4.1 General Rules for Test Assembly

The following are general rules, intended as guidelines for assembling test forms. When reviewing a test prior to administering, verify that the test conforms with the following test construction guidelines.

Test Construction Rules for Multiple-Choice Tests.

1. Set the number of items so that at least 95 percent of the examinees can answer all items.

2. The correct choice should appear about an equal number of times in each response position.

3. Do not use any pattern of correct responses, e.g., ABCDE, etc.

4. Directions to examinees should be written on the test to indicate whether guessing is permitted or not.

Test Construction Rules for Essay Tests.

1. All examinees must take the same items. Do not give them a chance to choose which items they want to answer. Meaningful comparisons normally can be made only if all examinees take the same test.

4.2 Grading Essay Tests

Because of their subjective nature, essay exams are difficult to grade. The following guidelines are helpful for grading essay exams in a consistent and meaningful way.

1. Construct a model answer for each item and award a point for each essential element of the model answer. This should help minimize the subjective effects of grading.

2. Essay items must be graded anonymously if at all possible in order to reduce the subjectivity of the graders. That is, graders should not be informed as to the identity of the examinees whose papers they are grading.

3. Grade a single essay item at a time. This helps the grader maintain a single set of criteria for awarding points to the response. In addition, it tends to reduce the influence of the examinee's previous performance on other items.

4. Unless it is a test of language mechanics, do not take off credit for poor handwriting, spelling errors, poor grammar, failure to punctuate properly, etc.

5. Ideally, there should be two graders for each item. Any disagreements between these two graders must be resolved by a third grader. Normally, this third grader is the head grader or course instructor.