Writing Multiple-Choice Test Items

Jerard Kehoe
Virginia Polytechnic Institute and State University

A notable concern of many teachers is that they frequently have the task of constructing tests but have relatively little training or information to rely on in this task. The objective of this Digest is to set out some conventional wisdom for the construction of multiple-choice tests, which are one of the most common forms of teacher-constructed tests. The comments which follow are applicable mainly to multiple-choice tests covering fairly broad topic areas.

Before proceeding, it will be useful to establish our terms for discussing multiple-choice items. The stem is the introductory question or incomplete statement at the beginning of each item and this is followed by the options. The options consist of the answer -- the correct option -- and distractors--the incorrect but (we hope) tempting options.

General Objectives
As a rule, one is concerned with writing stems that are clear and parsimonious, answers that are unequivocal and chosen by the students who do best on the test, and distractors that are plausible competitors of the answer as evidenced by the frequency with which they are chosen. Lastly, and probably most important, we should adopt the attitude that items need to be developed over time in the light of evidence that can be obtained from the statistical output typically provided by a measurement services office (where tests are machine-scored) and from "expert" editorial review.

Planning
The primary objective in planning a test is to outline the actual course content that the test will cover. A convenient way of accomplishing this is to take 10 minutes following each class to list on an index card the important concepts covered in class and in assigned reading for that day. These cards can then be used later as a source of test items. An even more conscientious approach, of course, would be to construct the test items themselves after each class. The advantage of either of these approaches is that the resulting test is likely to be a better representation of course activity than if the test were constructed before the course or after the course, when we usually have only a fond memory or optimistic syllabus to draw from. When we are satisfied that we have an accurate description of the content areas, then all that remains is to construct items that represent specific content areas. In developing good multiple-choice items, three tasks need to be considered: writing stems, writing options, and ongoing item development. The first two are discussed in this Digest.

Writing Stems
We will first describe some basic rules for the construction of multiple-choice stems, because they are typically, though not necessarily, written before the options.
1. Before writing the stem, identify the one point to be tested by that item. In general, the stem should not pose more than one problem, although the solution to that problem may require more than one step.
2. Construct the stem to be either an incomplete statement or a direct question, avoiding stereotyped phraseology, as rote responses are usually based on verbal stereotypes. For example, the following stems (with answers in parentheses) illustrate undesirable phraseology:

What is the biological theory of recapitulation? (Ontogeny repeats phylogeny)
Who was the chief spokesman for the "American System?" (Henry Clay)

Correctly answering these questions likely depends less on understanding than on recognizing familiar phraseology.
3. Avoid including nonfunctional words that do not contribute to the basis for choosing among the options. Often an introductory statement is included to enhance the appropriateness or significance of an item but does not affect the meaning of the problem in the item. Generally, such superfluous phrases should be excluded. For example, consider:

*The American flag has three colors. One of them is (1) red (2) green (3) black*

versus

*One of the colors of the American flag is (1) red (2) green (3) black*

In particular, irrelevant material should not be used to make the answer less obvious. This tends to place too much importance on reading comprehension as a determiner of the correct option.

4. Include as much information in the stem and as little in the options as possible. For example, if the point of an item is to associate a term with its definition, the preferred format would be to present the definition in the stem and several terms as options rather than to present the term in the stem and several definitions as options.

5. Restrict the use of negatives in the stem. Negatives in the stem usually require that the answer be a false statement. Because students are likely in the habit of searching for true statements, this may introduce an unwanted bias.

6. Avoid irrelevant clues to the correct option. Grammatical construction, for example, may lead students to reject options which are grammatically incorrect as the stem is stated. Perhaps more common and subtle, though, is the problem of common elements in the stem and in the answer. Consider the following item:

*What led to the formation of the States' Rights Party?*

- a. The level of federal taxation
- b. The demand of states for the right to make their own laws
- c. The industrialization of the South
- d. The corruption of federal legislators on the issue of state taxation

One does not need to know U.S. history in order to be attracted to the answer, b.

Other rules that we might list are generally commonsensical, including recommendations for independent and important items and prohibitions against complex, imprecise wording.

**Writing Options**

Following the construction of the item stem, the likely more difficult task of generating options presents itself. The rules we list below are not likely to simplify this task as much as they are intended to guide our creative efforts.

1. Be satisfied with three or four well constructed options. Generally, the minimal improvement to the item due to that hard-to-come-by fifth option is not worth the effort to construct it. Indeed, all else the same, a test of 10 items each with four options is likely a better test than a test with nine items of five options each.

2. Construct distractors that are comparable in length, complexity and grammatical form to the answer, avoiding the use of such words as "always," "never," and "all." Adherence to this rule avoids some of the more common sources of biased cueing. For example, we sometimes find ourselves increasing the length and specificity of the answer (relative to distractors) in order to insure its truthfulness. This, however, becomes an easy-to-spot clue for the testwise student. Related to this issue is the question of whether or not test writers should take advantage of these types of cues to construct more tempting distractors. Surely not! The number of students choosing a distractor should depend only on deficits in the content area which the item targets and should not depend on cue biases or reading comprehension differences in "favor" of the distractor.
3. Options which read "none of the above," "both a. and e. above," "all of the above," etc., should be avoided when the students have been instructed to choose "the best answer," which implies that the options vary in degree of correctness. On the other hand, "none of the above" is acceptable if the question is factual and is probably desirable if computation yields the answer. "All of the above" is never desirable, as one recognized distractor eliminates it and two recognized answers identify it.

4. After the options are written, vary the location of the answer on as random a basis as possible. A convenient method is to flip two (or three) coins at a time where each possible Head-Tail combination is associated with a particular location for the answer. Furthermore, if the test writer is conscientious enough to randomize the answer locations, students should be informed that the locations are randomized. (Testwise students know that for some instructors the first option is rarely the answer.)

5. If possible, have a colleague with expertise in the content area of the exam review the items for possible ambiguities, redundancies or other structural difficulties. Having completed the items we are typically so relieved that we may be tempted to regard the task as completed and each item in its final and permanent form. Yet, another source of item and test improvement is available to us, namely, statistical analyses of student responses.

This Digest was adapted with permission from Testing Memo 4: Constructing Multiple-Choice Tests -- Part I, Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

Further Reading

How to write Multiple Choice Questions
These topics below were prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education. You may distribute these documents freely.

1. Writing Multiple-Choice Test Items
2. More Multiple-choice Item Writing Do's And Don'ts
3. Making the A: How To Study for Tests Basic Item
4. Analysis for Multiple-choice Tests
5. Blooms Taxonomy

Writing Multiple-Choice Test Items
A notable concern of many teachers is that they frequently have the task of constructing tests but have relatively little training or information to rely on in this task. The objective of this Digest is to set out some conventional wisdom for the construction of multiple-choice tests, which are one of the most common forms of teacher-constructed tests. The comments which follow are applicable mainly to multiple-choice tests covering fairly broad topic areas.

Before proceeding, it will be useful to establish our terms for discussing multiple-choice items. The stem is the introductory question or incomplete statement at the beginning of each item and this is followed by the options. The options consist of the answer -- the correct option -- and distractors--the incorrect but (we hope) tempting options.

**General Objectives**

As a rule, one is concerned with writing stems that are clear and parsimonious, answers that are unequivocal and chosen by the students who do best on the test, and distractors that are plausible competitors of the answer as evidenced by the frequency with which they are chosen. Lastly, and probably most important, we should adopt the attitude that items need to be developed over time in the light of evidence that can be obtained from the statistical output typically provided by a measurement services office (where tests are machine-scored) and from "expert" editorial review.

**Planning**

The primary objective in planning a test is to outline the actual course content that the test will cover. A convenient way of accomplishing this is to take 10 minutes following each class to list on an index card the important concepts covered in class and in assigned reading for that day. These cards can then be used later as a source of test items. An even more conscientious approach, of course, would be to construct the test items themselves after each class. The advantage of either of these approaches is that the resulting test is likely to be a better representation of course activity than if the test were constructed before the course or after the course, when we usually have only a fond memory or optimistic syllabus to draw from. When we are satisfied that we have an accurate description of the content areas, then all that remains is to construct items that represent specific content areas. In developing good multiple-choice items, three tasks need to be considered: writing stems, writing options, and ongoing item development. The first two are discussed in this Digest.

**Writing Stems**

We will first describe some basic rules for the construction of multiple-choice stems, because they are typically, though not necessarily, written before the options.

1. Before writing the stem, identify the one point to be tested by that item. In general, the stem should not pose more than one problem, although the solution to that problem may require more than one step.

2. Construct the stem to be either an incomplete statement or a direct question, avoiding stereotyped phraseology, as rote responses are usually based on verbal stereotypes. For example, the following stems (with answers in parentheses) illustrate undesirable phraseology:
What is the biological theory of recapitulation? (Ontogeny repeats phylogeny)

Who was the chief spokesman for the "American System?” (Henry Clay)

Correctly answering these questions likely depends less on understanding than on recognizing familiar phraseology.

3. Avoid including nonfunctional words that do not contribute to the basis for choosing among the options. Often an introductory statement is included to enhance the appropriateness or significance of an item but does not affect the meaning of the problem in the item. Generally, such superfluous phrases should be excluded. For example, consider:

*The American flag has three colors. One of them is (1) red (2) green (3) black*

*versus*

*One of the colors of the American flag is (1) red (2) green (3) black*

In particular, irrelevant material should not be used to make the answer less obvious. This tends to place too much importance on reading comprehension as a determiner of the correct option.

4. Include as much information in the stem and as little in the options as possible. For example, if the point of an item is to associate a term with its definition, the preferred format would be to present the definition in the stem and several terms as options rather than to present the term in the stem and several definitions as options.

5. Restrict the use of negatives in the stem. Negatives in the stem usually require that the answer be a false statement. Because students are likely in the habit of searching for true statements, this may introduce an unwanted bias.

6. Avoid irrelevant clues to the correct option. Grammatical construction, for example, may lead students to reject options which are grammatically incorrect as the stem is stated. Perhaps more common and subtle, though, is the problem of common elements in the stem and in the answer. Consider the following item:

*What led to the formation of the States’ Rights Party?*

a. The level of federal taxation  
b. The demand of states for the right to make their own laws  
c. The industrialization of the South  
d. The corruption of federal legislators on the issue of state taxation

One does not need to know U.S. history in order to be attracted to the answer, b.

Other rules that we might list are generally commonsensical, including recommendations for independent and important items and prohibitions against complex, imprecise wording.

Writing Options
Following the construction of the item stem, the likely more difficult task of generating options presents itself. The rules we list below are not likely to simplify this task as much as they are intended to guide our creative efforts.

1. Be satisfied with three or four well constructed options. Generally, the minimal improvement to the item due to that hard-to-come-by fifth option is not worth the effort to construct it. Indeed, all else the same, a test of 10 items each with four options is likely a better test than a test with nine items of five options each.

2. Construct distractors that are comparable in length, complexity and grammatical form to the answer, avoiding the use of such words as "always," "never," and "all." Adherence to this rule avoids some of the more common sources of biased cueing. For example, we sometimes find ourselves increasing the length and specificity of the answer (relative to distractors) in order to insure its truthfulness. This, however, becomes an easy-to-spot clue for the testwise student. Related to this issue is the question of whether or not test writers should take advantage of these types of cues to construct more tempting distractors. Surely not! The number of students choosing a distractor should depend only on deficits in the content area which the item targets and should not depend on cue biases or reading comprehension differences in "favor" of the distractor.

3. Options which read "none of the above," "both a. and e. above," "all of the above," ... should be avoided when the students have been instructed to choose "the best answer," which implies that the options vary in degree of correctness. On the other hand, "none of the above" is acceptable if the question is factual and is probably desirable if computation yields the answer. "All of the above" is never desirable, as one recognized distractor eliminates it and two recognized answers identify it.

4. After the options are written, vary the location of the answer on as random a basis as possible. A convenient method is to flip two (or three) coins at a time where each possible Head-Tail combination is associated with a particular location for the answer. Furthermore, if the test writer is conscientious enough to randomize the answer locations, students should be informed that the locations are randomized. (Testwise students know that for some instructors the first option is rarely the answer.)

5. If possible, have a colleague with expertise in the content area of the exam review the items for possible ambiguities, redundancies or other structural difficulties. Having completed the items we are typically so relieved that we may be tempted to regard the task as completed and each item in its final and permanent form. Yet, another source of item and test improvement is available to us, namely, statistical analyses of student responses.

This Digest was adapted with permission from Testing Memo 4: Constructing Multiple-Choice Tests -- Part I, Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

Further Reading


More Multiple-choice Item Writing

Do's And Don'ts

Robert B. Frary

Virginia Polytechnic Institute and State University

ERIC Digest TM 95-3 gave a few suggestions for item-writing, but only to a limited extent, due to its coverage of other aspects of test development. What follows here is a fairly comprehensive list of recommendations for writing multiple choice items. Some of these are backed up by psychometric research; i.e., it has been found that, generally, the resulting scores are more accurate indicators of each student's knowledge when the recommendations are followed than when they are violated. Other recommendations result from logical deduction.

Content
1. Do ask questions that require more than knowledge of facts. For example, a question might require selection of the best answer when all of the options contain elements of correctness. Such questions tend to be more difficult and discriminating than questions that merely ask for a fact. Justifying the "bestness" of the keyed option may be as challenging to the instructor as the item was to the students, but, after all, isn't challenging students and responding to their challenges a big part of what being a teacher is all about?
2. Don't offer superfluous information as an introduction to a question, for example, "The presence and association of the male seems to have profound effects on female physiology in domestic animals. Research has shown that in cattle presence of a bull has the following effect:"
   This approach probably represents an unconscious effort to continue teaching while testing and is not likely to be appreciated by the students, who would prefer direct questions and less to read. The stem just quoted could be condensed to "Research has shown that the presence of a bull has which of the following effects on cows?" (17 words versus 30).

Structure
3. Don't ask a question that begins, "Which of the following is true [or false]?" followed by a collection of unrelated options. Each test question should focus on some specific aspect of the course. Therefore, it's OK to use items that begin, "Which of the following is true [or false] concerning X?" followed by options all pertaining to X. However, this construction should be used sparingly if there is a tendency to resort to trivial reasons for falseness or an opposite tendency to offer options that are too obviously true. A few true-false questions (in among the multiple-choice questions) may forestall these problems. The options would be: 1) True 2) False.
4. Don't use items like the following:

   What is (are) the capital(s) of Bolivia?

   A. La Paz B. Sucre C. Santa Cruz

   1) A only 4) Both A and B
   2) B only 5) All of the above
   3) C only

   Research on this item type has consistently shown it to be easier and less discriminating than items with distinct options. In the example above, one only needs to remember that Bolivia has
two capitals to be assured of answering correctly. This problem can be alleviated by offering all possible combinations of the three basic options, namely:

1) A only, 2) B only, 3) C only, 4) A and B, 5) A and C, 6) B and C, 7) A, B, and C, 8) None of the above.

However, due to its complexity, initial use of this adaptation should be limited.

**Options**

5. Do ask questions with varying numbers of options. There is no psychometric advantage to having a uniform number, especially if doing so results in options that are so implausible that no one or almost no one marks them. In fact, some valid and important questions demand only two or three options, e.g., "If drug X is administered, body temperature will probably: 1) increase, 2) stay about the same, 3) decrease."

5. Don't put negative options following a negative stem. Empirically (or statistically) such items may appear to perform adequately, but this is probably only because brighter students who naturally tend to get higher scores are also better able to cope with the logical complexity of a double negative.

7. Don't use "all of the above." Recognition of one wrong option eliminates "all of the above," and recognition of two right options identifies it as the answer, even if the other options are completely unknown to the student. Probably some instructors use items with "all of the above" as yet another way of extending heir teaching into the test (see 2 above). It just seems so good to have the students affirm, say, all of the major causes of some phenomenon. With this approach, "all of the above" is the answer to almost every item containing it, and the students soon figure this out.

8. Do ask questions with "none of the above" as the final option, especially if the answer requires computation. Its use makes the question harder and more discriminating, because the uncertain student cannot focus on a set of options that must contain the answer. Of course, "none of the above" cannot be used if the question requires selection of the best answer and should not be used following a negative stem. Also, it is important that "none of the above" should be the answer to a reasonable proportion of the questions containing it.

9. Don't include superfluous information in the options. The reasons given for 8 above apply. In addition, as another manifestation of the desire to teach while testing, the additional information is likely to appear on the correct answer: 1) W, 2) X, 3) Y, because ..., 4) Z. Students are very sensitive to this tendency and take advantage of it.

10. Don't use specific determiners in distractors. Sometimes in a desperate effort to produce another, often unneeded, distractor (see 5 above), a statement is made incorrect by the inclusion of words like all or never, e.g., "All humans have 46 chromosomes." Students learn to classify such statements as distractors when otherwise ignorant.

11. Don't repeat wording from the stem in the correct option. Again, an ignorant student will take advantage of this practice.

**Errors to avoid**

Most violations of the recommendations given thus far should not be classified as outright errors, but, instead, perhaps, as lapses of judgement. And, as almost all rules have exceptions, there are probably circumstances where some of 1-11 above would not hold. However, there are three not-too-common item-writing/test-preparation errors that represent nothing less than negligence. They are now mentioned to encourage careful preparation and proofreading of tests:

**Typos.** These are more likely to appear in distractors than in the stem and the correct answer, which get more scrutiny from the test preparer. Students easily become aware of this tendency if it is present.

**Grammatical inconsistency** between stem and options. Almost always, the stem and the correct answer are grammatically consistent, but distractors, often produced as afterthoughts, may not mesh properly with the stem. Again, students quickly learn to take advantage of this foible.

**Overlapping distractors.** For example: *Due to budget cutbacks, the university library now subscribes to fewer than _?_ periodicals. 1) 25,000 2) 20,000 3) 15,000 4) 10,000*

Perhaps surprisingly, not all students "catch on" to items like this, but many do. Worse yet, the instructor might indicate option 2 as the correct answer.

Finally, we consider an item-writing foible reported by Smith (1982). What option would you select among the following (stem omitted)?
1) Abraham Lincoln 3) Stephen A. Douglas
2) Robert E. Lee 4) Andrew Jackson

The testwise but ignorant student will select Lincoln because it represents the intersection of two categories of prominent nineteenth century people, namely, presidents and men associated with the Civil War. This one:

before breakfast 3) on a full stomach

ith meals 4) before going to bed

Two options have to do with eating, and two with the time of day. Only one relates to both. Unfortunately, item writers consciously or unconsciously construct items of this type with the intersection invariably the correct answer.

This digest was adapted with permission from Testing Memo 10: Some Multiple-choice Item Writing Do's And Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060.

Reading

important information to study. Reviewing regularly allows you to avoid cramming and reduces test anxiety. The biggest benefit is it gives you time to absorb information.

Read difficult assignments twice. Sometimes a second reading will clarify concepts. If you are having difficulty with a subject, get help immediately. Meet with your instructor after class, use an alternate text to supplement required reading or hire a tutor (ask faculty members and other students for referrals).

**Review, Review, Review**

Plan ahead, scheduling review periods well in advance. Set aside one hour on a Saturday or Sunday to review several subjects. Keep your reviews short and do them often.

- **Daily reviews**—Conduct short before and after class reviews of lecture notes. Begin reviewing after your first day of class.

- **Weekly reviews**—Dedicate about 1 hour per subject to review assigned reading and lecture notes.

- **Major reviews**—Start the week before an exam and study the most difficult subjects when you are the most alert. Study for 2-5 hours punctuated by sufficient breaks.

Create review tools, such as flashcards, chapter outlines and summaries. This helps you organize and remember information as well as condense material to a manageable size. Use 3 x 5 cards to review important information. Write ideas, formulas, concepts and facts on cards to carry with you. Study on the bus, in waiting rooms or whenever you have a few extra minutes.

Another useful tool is a study checklist. Make a list of everything you need to know for the exam. The list should include a brief description of reading assignments, types of problems to solve, skills to master, major ideas, theories, definitions, and equations. When you begin your final study sessions, cross off items as you review them.

**Study Groups**

For some subjects, study groups are an effective tool. Study groups allow students to combine resources; members share an academic goal and provide support and encouragement. Such groups meet regularly to study and learn a specific subject.

To form a study group, look for dedicated students—students who ask and answer questions in class, and who take notes. Suggest to two or three that you meet to talk about group goals, meeting times and other logistics. Effective study groups are limited to five or six people. Test the group first by planning a one-time-only session. If that works, plan another. After several successful sessions, schedule regular meetings.

Set an agenda for each meeting to avoid wasting time. List the material that will be reviewed so members can come prepared. Also, follow a format. For example, begin by comparing notes to make sure you all heard the same thing and recorded important information. Spend 15-20 minutes conducting open-ended discussions on specific topics. Then, test each other by asking questions or take turns explaining concepts. Set aside 5-10 minutes to brainstorm possible test questions.
Taking an Exam

On exam day arrive early and get organized. Pay attention to verbal directions as tests are distributed. Read directions slowly. Scan the entire test, noticing how many points each part is worth and estimate the time needed for individual questions. Before you start answering questions, write down memory aids, formulas, equations, facts and other useful information in the margins.

Check the time and pace yourself. If you get stuck on a question try to remember a related fact. Start from the general and go to the specific. Look for answers in other test questions. Often a term, name, date or other fact you have forgotten will appear somewhere else in the test. Move on to the next question if memory aids do not help. You can always go back to the question if you have time.

Test-Taking Tips for Different Types of Exams

Multiple Choice--Check the directions to see if the questions call for more than one answer. Answer each question in your head before you look at the possible answers. If you can come up with the answer before you look at the choices you eliminate the possibility of being confused by them. Mark questions you can't answer immediately and come back to them later.

When taking a multiple-choice exam guess only if you are not penalized for incorrect answers. Use the following guidelines to make educated guesses.

- If two answers are similar, except for one or two words, choose one of these answers.
- If the answer calls for a sentence completion, eliminate the answers that would not form grammatically correct sentences.
- If answers cover a wide range (5, 76, 87, 109, 500) choose a number in the middle.

For machine-graded multiple-choice tests be certain that the answer you mark corresponds to the question you are answering. Check the test booklet against the answer sheet whenever you start a new section and again at the top of each column.

True-false--If any part of a true-false statement is false, the answer is false. Look for key words, i.e., qualifiers like all, most, sometimes, never or rarely. Questions containing absolute qualifiers such as always or never often are false.

Open book--When studying for this type of test, write down any formulas you will need on a separate sheet. Place tabs on important pages of the book so that you don't have to waste time looking for tables or other critical information. If you plan to use your notes, number them and make a table of contents. Prepare thoroughly for open-book tests. They are often the most difficult.

Short answer/fill-in-the-blank--These tests require students to provide definitions or short descriptions (typically a few words or a sentence or two). Study using flashcards with important terms and phrases. Key words and facts will then be familiar and easy to remember as you answer test questions.
Essay—When answering an essay question, first decide precisely what the question is asking. If a question asks you to compare, do not explain. Standard essay question words are listed in Box 1. Look up any unfamiliar words in a dictionary.

Before you write your essay, make a quick outline. There are three reasons for doing this. First, your thoughts will be more organized (making it easier for your teacher to read), and you will be less likely to leave out important facts. Second, you will be able to write faster. Third, if you do not have time to finish your answer, you may earn some points with the outline. Don't forget to leave plenty of space between answers. You can use the extra space to add information if there is time.

**Verbs Commonly Used in Essay Questions**

- Analyze
- Compare
- Contrast
- Criticize
- Define
- Describe
- Discuss
- Enumerate
- Evaluate
- Examine
- Explain
- Illustrate
- Interpret
- List
- Outline
- Prove
- State
- Summarize

When you write, get to the point. Start off by including part of the question in your answer. For example, if the question asks, "Discuss the benefits and drawbacks of universal health care coverage to both patients and medical professionals." Your first sentence might read, "Universal health care will benefit patients in the following ways." Expand your answer with supporting ideas and facts. If you have time, review your answers for grammatical errors, clarity and legibility.

**Further Reading**


Basic Item Analysis for Multiple-choice Tests

Jerard Kehoe
Virginia Polytechnic Institute and State University

This Digest offers some suggestions for the improvement of multiple-choice tests using "item analysis" statistics. These statistics are typically provided by a measurement services, where tests are machine-scored, as well as by testing software packages.

The basic idea that we can capitalize on is that the statistical behavior of "bad" items is fundamentally different from that of "good" items. Of course, the items have to be administered to students in order to obtain the needed statistics. This fact underscores our point of view that tests can be improved by maintaining and developing a pool of "good" items from which future tests will be drawn in part or in whole. This is particularly true for instructors who teach the same course more than once.

What Makes an Item Psychometrically Good?

In answering this question, it is desirable to restrict our discussion to tests which are written to cover a unified portion of course material such that it is unlikely that a student would do well on one part of a test and poorly on another. If this latter situation is the case, the comments which follow will apply only if the corresponding topics are tested separately. Regardless, this approach would be preferred, because, otherwise, scores would be ambiguous in their reporting of students' achievement.

Once the instructor is satisfied that the test items meet the above criterion and that they are indeed appropriately written, what remains is to evaluate the extent to which they discriminate among students. The degree to which this goal is attained is the basic measure of item quality for almost all multiple-choice tests. For each item the primary indicator of its power to discriminate students is the correlation coefficient reflecting the tendency of students selecting the correct answer to have high scores. This coefficient is reported by typical item analysis programs as the item discrimination coefficient or, equivalently, as the point-biserial correlation between item score and total score. This coefficient should be positive, indicating that students answering correctly tend to have higher scores. Similar coefficients may be provided for the wrong choices. These should be negative, which means that students selecting these choices tend to have lower scores.

Alternatively, some item analysis programs provide the percentages of examinees scoring in the top, middle, and bottom thirds who select each option. In this case, one would hope to find that large proportions of the high scorers answered correctly, while larger proportions of low scorers selected the distractors.

The proportion of students answering an item correctly also affects its discrimination power. This point may be summarized by saying that items answered correctly (or incorrectly) by a large proportion of examinees (more than 85%) have markedly reduced power to discriminate. On a good test, most items will be answered correctly by 30% to 80% of the examinees.
A general indicator of test quality is the reliability estimate usually reported on
the test scoring/analysis printout. Referred to as KR-20 or Coefficient Alpha, it
reflects the extent to which the test would yield the same ranking of examinees if
readministered with no effect from the first administration, in other words, its
accuracy or power of discrimination. Values of as low as .5 are satisfactory for
short tests (10 - 15 items), though tests with over 50 items should yield KR-20
values of .8 or higher (1.0 is the maximum). In any event, important decisions
concerning individual students should not be based on a single test score when the
Corresponding KR-20 is less than .8. Unsatisfactorily low KR-20s are usually due
to an excess of very easy (or hard) items, poorly written items that do not
discriminate, or violation of the precondition that the items test a unified body of
content.

Improving the ability of items to discriminate.

The statistics usually provided by a test scoring service provide the information
needed to keep a record of each item with respect to its performance. One
approach is simply to tape a copy of each item on a 5 x 7 card with the test
content area briefly described at the top. In addition, tape the corresponding line
from the computer printout for that item each time it is used. Alternatively, item
banking programs may provide for inclusion of the proportions marking each
option and item discrimination coefficients along with each item's content.

A few basic rules for item development follow:

1. Items that correlate less than .15 with total test score should probably be
restructured. One's best guess is that such items do not measure the same
skill or ability as does the test on the whole or that they are confusing or
misleading to examinees. Generally, a test is better (i.e., more reliable) the
more homogeneous the items. Just how to restructure the item depends
largely on careful thinking at this level. Begin by applying the rules of
stem and option construction discussed in ERIC Digest TM 95-2. If there
are any apparent violations, correct them on the 5x7 card or in the item
bank. Otherwise, it's probably best to write a new item altogether after
considering whether the content of the item is similar to the content
objectives of the test.

2. Distractors that are not chosen by any examinees should be replaced or
eliminated. They are not contributing to the test's ability to discriminate
the good students from the poor students. One should not be concerned if
each distractor is not chosen by the same number of examinees. Different
kinds of mistakes may very well be made by different numbers of
students. Also, the fact that a majority of students miss an item does not
imply that the item should be changed, although such items should be
double-checked for their accuracy. One should be suspicious about the
correctness of any item in which a single distractor is chosen more often
than all other options, including the answer, and especially so if that
distractor's correlation with the total score is positive.

3. Items that virtually everyone gets right are useless for discriminating
among students and should be replaced by more difficult items. This
recommendation is particularly true if you adopt the traditional attitude
toward letter grade assignments that letter grades more or less fit a predetermined distribution.

By constructing, recording, and adjusting items in this fashion, teachers can develop a pool of items for specific content areas with conveniently available resources.

**Some Further Issues.**

The suggestions here focus on the development of tests which are homogeneous, that is, tests intended to measure a unified content area. Only for such tests is it reasonable to maximize item-test correlations or, equivalently, KR-20 or Coefficient Alpha (reliability), which is the objective of step 1 above. The extent to which a high average item-test correlation can be achieved depends to some extent on the content area.

It is generally acknowledged that well constructed tests in vocabulary or mathematics are more homogeneous than well constructed tests in social sciences. This circumstance suggests that particular content areas have optimal levels of homogeneity and that these vary from discipline to discipline. Perhaps psychologists should strive for lower test homogeneity than mathematicians because course content is less homogeneous.

A second issue involving test homogeneity is that of the precision of a student's obtained test score as an estimate of that student's "true" score on the skill tested. Precision (reliability) increases as the average item-test correlation increases, all else the same; and precision decreases as the number of items decreases, all else the same.

These two relationships lead to an interesting paradox: often the precision of a test can be increased simply by discarding the items with low item-test correlations. For example, a 30-item multiple-choice test administered by the author resulted in a reliability of .79, and discarding the seven items with item-test correlations below .20 yielded a 23-item test with a reliability of .88! That is, by dropping the worst items from the test, the students' obtained scores on the shorter version are judged to be more precise estimates than the same students' obtained scores on the longer version.

The reader may question whether it is ethical to throw out poorly performing questions when some students may have answered them correctly based on their knowledge of course material. Our opinion is that this practice is completely justified. The purpose of testing is to determine each student's rank. Retaining psychometrically unsatisfactory questions is contrary to this goal and degrades the accuracy of the resulting ranking.

This Digest was adapted with permission from *Testing Memo 5: Constructing Multiple-Choice Tests -- Part II*, Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

**Further Reading**


From Bloom, et al., 1956

As teachers we tend to ask questions in the "knowledge" category 80% to 90% of the time. These questions are not bad, but using them all the time is. Try to utilize higher order level of questions. These questions require much more "brain power" and a more extensive and elaborate answer. Below are the six question categories as defined by Bloom.

**• KNOWLEDGE**
- remembering;
- memorizing;
- recognizing;
- recalling identification and recall of information
  - Who, what, when, where, how ...?
  - Describe

**• COMPREHENSION**
- interpreting;
- translating from one medium to another;
- describing in one's own words;
- organization and selection of facts and ideas
  - Retell...

**• APPLICATION**
- problem solving;
- applying information to produce some result;
- use of facts, rules and principles
  - How is...an example of...?
  - How is...related to...?
• Why is...significant?

• ANALYSIS
  • subdividing something to show how it is put together;
  • finding the underlying structure of a communication;
  • identifying motives;
  • separation of a whole into component parts
    • What are the parts or features of...?
    • Classify...according to...
    • Outline/diagram...
    • How does...compare/contrast with...?
    • What evidence can you list for...?

• SYNTHESIS
  • creating a unique, original product that may be in verbal form or may be a physical object;
  • combination of ideas to form a new whole
    • What would you predict/infer from...?
    • What ideas can you add to...?
    • How would you create/design a new...?
    • What might happen if you combined...?
    • What solutions would you suggest for...?

• EVALUATION
  • making value decisions about issues;
  • resolving controversies or differences of opinion;
  • development of opinions, judgements or decisions
    • Do you agree...?
    • What do you think about...?
    • What is the most important...?
    • Place the following in order of priority...
    • How would you decide about...?
• What criteria would you use to assess...?