For the Teacher . . .

How to Develop Multiple-Choice Tests

Lee Sechrest, John F. Kihlstrom, and
Richard R. Bootzin
University of Arizona

Sooner or later, almost everyone who teaches psychology has to develop a multiple-choice test. Many lower-level textbooks come with sets of questions distributed by the publishers, but those items do not cover unique lecture material. Moreover, even the published items ought to be carefully reviewed, for they may not be of uniformly high quality. And, many of them violate what might be considered standard rules for writing good items (Ellsworth, Dunnell, and Duell, 1990). In any case, beyond lower-level courses, teachers are likely to be mostly on their own with respect to developing multiple-choice tests. What is one to do?

Start at the Beginning

Why give a test in the first place? The easy answer is, "Well, of course, we want to find out how much students have learned." The difficulty is we can't know how much students have learned without first having a pretest or baseline measure. We can't assume students have started at zero (or at any specific level greater than chance). Houston (1985) gave a set of test items from an introduction to psychology to a group of 60 diverse persons (aged 16-61) with no formal training in psychology, and this group exceeded chance performance on 76 percent of the items.

Start at the End

The temptation is to propose a "final" exam at the *outset* of the course and to consider this baseline in evaluating performance on subsequent tests, but that's got its own set of problems. So, an alternative answer to the question "Why give a test...?" is that we at least want to know how much the examinees know about psychology. In that case, we could identify some more or less critical knowledge—that students ought to have acquired—and determine how many actually did acquire that knowledge. That would involve the construction of a *criterion-referenced* test. That is, specify a criterion level that represents successful performance.

If you choose this approach, you are not much concerned about differences *among* students above or below the criterion level. For example, if teaching differential equations, you're not much interested in the fact that some inept students were more inept than others or that some students could do the problems faster—but no more correctly—than others. Or, if teaching psychology using the Keller (1968) method—in which all students are expected to reach criterion performance—there would be little interest in differences among students.

Norm-referenced Tests

Usually, however, teachers construct exams to distinguish between students with different levels of knowledge—in other words, they create *norm-referenced* tests. They want to order students with respect to their knowledge, and to be able to say—for any level of performance—which students know more. That aim raises issues (i.e., reliability and homogeneity of sets of items) that are not inherent in criterion-referenced measures and that should affect how items are written and scored.

Reliability, in this context, refers to the dependability of conclusions about differences in ability inferred from test scores. Nearly everyone understands that on, let us say, a 60-item test, the difference between 42 correct and 43 correct is trivial. (That is why most instructors like to position the borderlines between letter grades at points at which natural breaks in a distribution of scores occur. This minimizes the number of students whose letter grade is affected by a single answer.)

Half Wrong, Half Right: Uncovering a Difference

In order for test scores to allow some ordering among test takers, items must be constructed to result in a distribution of scores. That is, the variance of the score distribution should be large. The maximum variance for any given item will occur when its difficulty level is .50 (i.e., when half the respondents get it correct and half get it wrong).

In general, differentiation among examinees on a test will be greatest when the difficulty level of the test causes subjects to get about half the items correct. Put another way, neither very difficult nor very easy items assist much in differentiating among examinees.

Ideally, each and every item in a test should "work," that is, help differentiate maximally among examinees. Teachers may try to vary item difficulty by writing some easy and some difficult items in the mistaken belief that these test results will better represent the knowledge distribution. But reduced variance contributes to reduced differentiation.

Even so, there are good reasons to include some easy and difficult items. Some teachers reason that including some easy and some difficult items may serve motivational purposes. Easy items may reduce anxiety about the exam and difficult items may reassure the best students that their knowledge is being fairly evaluated. Difficult items may also serve a diagnostic function for the instructor who may want to know whether a specific construct has been learned by the students who have best mastered the material.

Difficulty Level and Fairness

Items with a difficulty level of .50 are not easy to write, and a test consisting only of such items may be somewhat demoralizing to students expecting to do better than 50 percent correct. Item variance is actually not much reduced unless the difficulty level is fairly extreme, say beyond .80. That is, if the correct/incorrect split on items is not worse than .80/.20 (or .20/.80), variance is not much reduced. For the sake of student morale, a test with a mean percentage correct of about 70 may be desirable.

Also, a test should be fair. Students do sometimes complain that a particular item is too detailed, ambiguous, or otherwise inappropriate. From a psychometric point of view there is a clear and objective index of fairness: the item-to-total correlation. An item belongs on a test if the correct response is positively correlated with scores on the remainder of the test. Most test-scoring software has the capacity to calculate these correlations. Of course, with a large N (such as that encountered in most introductory and survey courses) even very small correlations become statistically significant. A reasonable threshold for retaining an item might be that its item-to-total correlation should be at least .20 (for $N = 100$, this correlation is significant at $p < .05$); items failing to meet that criterion then would be eliminated from the test (e.g., by scoring the item correct for all responses).

Students immediately grasp the idea behind this practice and appreciate the extra effort entailed in rescoring the test to ensure fairness. And when confronted with the fact that a particular item did in fact discriminate between high and low scorers on the test, their complaints are almost always withdrawn. This assumes, of course, that most of the items on the test are perceived as fair. It is unlikely, but possible, to construct entire tests in which the variability between students is due to irrelevant considerations rather than to knowledge about the course material. In those cases, item-to-total correlations are not helpful.

Sources of Variance

Variance in test scores is determined in complex ways. Preferably, nearly all the variance should be determined by differences in knowledge at the time the test is given. In fact, however, the number and variety of determinants of variance will be large. Scores will vary because, among other reasons, some students: (1) are better and faster readers than others; (2) are test-wise (i.e., have learned heuristics to identify an answer that has a good chance of being right); (3) are smart enough or lucky enough to sit next to a better student from whose paper they may copy; (4) will have been lucky enough to have studied the exact material on which a few items are based; (5) will be relaxed and in a good frame of mind for taking the test, while others are anxious and distracted.

Instructors can fairly easily reduce some of these sources of variance (e.g., cheating, reading ability) but must accept other sources (e.g., luck in what was studied). Instructors should certainly construct items to minimize unwanted sources of variance. Correct response choices should, for example, be balanced across the options so that any position bias (e.g., the inclination to choose the last alternative) should not be either an advantage or a disadvantage. Characteristics of response alternatives not reflecting particular content (e.g., length, format) should not be cues to correctness. Extraneous material and difficult vocabulary should be excluded from the stems and distractors of items so that reading ability is minimized as a source of variance.

Research Basis of Item Construction

Advice about how to write multiple-choice items is not scarce. For example, a study of educational psychology textbooks found guidelines offered in 32 of 42 texts, and 12 of the guidelines were given in half or more of the 32 (Ellsworth, Dunnell, and Duell, 1990). Unfortunately, most of the advice is, apparently, just that—advice. Empirical support for many guidelines is lacking (Haladyna and Downing, 1989), but where it does exist, the support is usually thin, being limited to a study or two of dubious generalizability.

A Summary

Nonetheless, a review of empirical support, combined in an informal Bayesian way with expert opinion, reported by Haladyna and Downing (1989), is useful. (For a more general summary of research and expert opinion, see McKeachie, 1986.) We summarize, and edit, to some extent, their recommendations here:

1. Consider using only three instead of the usual four or five options for questions. Item statistics are generally as good with three options as with four or five, and because time per item is reduced, the number of items and content covered can be increased. Very often, it is difficult to come up with three or more good distractors anyway.
2. Balance the key so that the correct answer appears approximately equally often in every position. Students who have a tendency to choose one alternative (e.g., the last one) whenever they are uncertain should be neither more nor less likely to be right across items than would be expected by chance.
3. Do not use "all of the above," "none of the above," and similar alternatives as possible answers. Such choices generally make items a bit more difficult but are not helpful in other ways, since they introduce additional response biases. Also, do not use "I do not know" as a response option; after all, arguably, in many cases this answer is literally correct.
4. Keep lengths of options fairly consistent within items (e.g., so that the correct response is not notably longer than the distractors), and avoid giving the answer away by grammatical construction of the item. For example:

   An episcotister is an
   (a) instrument
   (b) computer software
   (c) theoretical construct

SEE TESTS ON PAGE 22

TESTS FROM PAGE 10

5. Try to use only plausible distractors and avoid distractors that contain clues that might be used by test-wise examinees. Ideally, for classroom tests, distractors should be diagnostic in the sense that incorrect answers should reveal specific deficits in knowledge or lapses in thinking. For example:

A dog hears a tone immediately before a puff of air is presented to the cornea of its eye. The puff of air is the:
(a) conditioned stimulus
(b) distal stimulus
(c) unconditioned stimulus
(d) generalization stimulus

An implausible distractor should attract very few responses, and thus represents a nonfunctional response option. Distractors including such adverbs as "never" and "always" tend to be avoided by test-wise students, when they are uncertain, and such distractors tend to produce biased patterns of responding that may favor one group of respondents over another. That is, the final score distribution will have an unwanted component of variance that is systematic but unrelated to knowledge of the material.

Distractors

The purpose of distractors is to reduce the probability that a student can get the correct answer to a question by guessing. For that to happen, distractors must attract a reasonable share of responses. An implausible or nonsense distractor, in effect, changes the difficulty level of an item. For example, in a four-choice item, the chance level of difficulty for the item is .25. But if one of the distractors is a throw-away, the chance level for the item becomes .33. That does not necessarily hurt either the reliability or the validity of the test, but it does change how you interpret how much students have learned.

Our personal experience suggests that tests should begin with three or four fairly easy items so that anxious students are not "paralyzed" immediately by difficult material. Other than that, ordering test items more or less in the sequence in which the material was presented in books and lectures seems to help students do better (Balch, 1989). To the extent that such an order effect is constant across students, it has no effect on variance, and, hence, on differentiating between students. It may, however, put students more at ease, particularly if the test is difficult.

Reduction of Irrelevant Variance

To reduce variance associated with individual differences in test-wiseness, as opposed to competence in the course, we also suggest informing all students, at the outset of the exam, of useful test-taking strategies. These include: (1) reading the test all the way through before answering any items (because one item may give hints about another); (2) trying to eliminate at least one option as clearly wrong (thereby increasing the likelihood of getting the item right by chance); (3) reasoning to the correct answer (when fact retrieval fails) from some general concept or principle (assuming that the instructor has not nefariously asked a question about an exception that tests the rule); and (4) guessing (when all else fails), because in the absence of explicit memory, implicit memory for studied material is likely to bias responding toward the correct answer. Instructors probably should not offer the disclaimer proposed by the public-radio humorist Michael Feldman: "All questions have been carefully researched, though the answers have not; ambiguous, misleading, and poorly worded questions are par for the course."

True or False?

Very often the idea to be tested by an item may lend itself better to a true-false than a multiple-choice format. For one thing, it may be difficult to come up with three or four good distractors. Besides, if the distractors are poor ones, the item may be inadvertently converted to a two-choice item (i.e., the equivalent of a true-false item). For example, the item: Who first formulated the concept of correlation:
(a) Karl Marx
(b) Ronald Fisher
(c) Francis Galton
(d) Sigmund Freud

would for informed students be a two-choice item that could be rephrased as "Ronald Fisher first formulated the concept of correlation: true or false?" If the answer is false, then the correct answer to the item must be Galton.

There is nothing wrong with true-false items; in fact, they result in tests with about the same psychometric properties as multiple-choice tests. True-false tests are likely to produce higher overall scores since chance-level performance is .5. What probably is not a good idea is mixing multiple-choice and true-false items in the same section of a test. Mixing item types tends to produce response errors that have nothing to do with what students know. Multiple-choice and true-false items used in the same test probably should be separated into two sections, preferably with answer sheets marked in such a way that the student cannot put a mark in a wrong space.

Objections to Multiple-Choice Tests?

The usual objection to multiple-choice tests is that they reflect only rather low-level memory processes rather than the higher-order concepts deemed "really important." However, there is no reason why multiple-choice tests cannot tap fairly abstract, conceptual knowledge. Consider the following item:

The fundamental process in classical conditioning is:
(a) association by contiguity
(b) vicarious reinforcement
(c) association by contingency
(d) continuous reinforcement

Now consider the following alternative:

In a classical conditioning experiment, a tone CS is paired with an electric shock US. For Group A, the CS precedes the US by 10 seconds. For Group B, the CS and US are presented simultaneously. For Group C, the US precedes the CS by 10 seconds. After 20 conditioning trials, the experimenter measures the magnitude of the fear CR. The most likely ordering of the CR magnitudes is:
(a) B > C = A
(b) B > A > C
(c) A > B > C
(d) A = C > B

Arguably, a student who gets the alternative item correct has a fairly good conceptual understanding of classical conditioning, at the level appropriate for Introductory Psychology.

If items designed to measure higher-order concepts correlate highly with items depending more clearly on memory (e.g., Ferland, Dorval, and Levasseur, 1987), is that an indictment of multiple-choice tests? Not necessarily, for the results suggest just as strongly that memory functions are related to those involved in higher order cognitive processes. Amazingly, after years of multiple-choice testing, we still do not have a very good notion of just what functions are tapped by such tests. In the meantime, we rely on the widely shared observation that it is unusual to find a student who does well on a multiple-choice test who is at the same time incapable of displaying other forms of comprehension of the course material.

Of course, this is an empirical question begging to be investigated. In general, we encourage teachers to experiment with tests. What exactly *is* the correlation between multiple-choice, short-answer, and essay tests of the same material? Does performance on items drawn from the text correlate with performance on items drawn from lectures? If a test is factor-analyzed, will the resulting structure mirror the organization of the course? When an instructor relocates to another institution, it may be useful for him or her to repeat readings, lectures, and exams from the previous year and to measure differences in student performance. This may yield useful clues about differences in the student populations being served.

Finally, tests are intended to evaluate, and promote, the learning process. Students should be encouraged to do more than score their tests against a key, count up the number correct, and slink away. Rather, they should be encouraged to treat the exam itself as a learning experience—to try to determine mastery of the course material. Instructors should consider preparing detailed feedback on their exams, perhaps short essays indicating what the question was about, why the right answer was right, and the wrong answers wrong. And, of course, similar considerations apply to the instructor. If students consistently do poorly on items testing particular concepts or principles, then the text or lecture material is a candidate for revision.

References

Balch, W.R. (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology*, 16, 75-77.

Ellsworth, R.A., Dunnell, P., and Duell, O.K. (1990). Multiple-choice test items: What are textbook authors telling teachers? *Journal of Educational Research*, 83, 289-293.

Ferland, J.J., Dorval, J., and Levasseur, L. (1987). Measuring higher cognitive levels by multiple choice questions: A myth? *Medical Education*, 21, 109-113.

Haladyna, T.M., and Downing, S.M. (1989). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2, 51-78.

Houston, J.P. (1985). Untutored lay knowledge of the principles of psychology: Do we know anything they don't? *Psychological Reports*, 70, 567-570.

Keller, F.S. (1968). RGood-bye teacher...S. *Journal of Applied Behavior Analysis*, 1, 79-89.

McKeachie, W.J. (1986). *Teaching tips: A guidebook for the beginning college teacher.* 8th Ed. Lexington, MA.: Heath.