TESTS & TESTING STANDARDS

Inside this document…

CONTENTS

INTRODUCTION TO TESTS

We Hate and Need Tests

It seems that tests are almost universally hated – or at least dreaded. Students groan when a parent or teacher announces an upcoming test. A teenager may have problems sleeping as he studies and practices for his driver's license test. A law student obsesses over passing her bar exam.

On the other hand, tests are critical to success. If my car is sputtering, I expect the mechanic to run some tests on it before replacing the engine. If I experience chest pains or have problems reading road signs, I know that I will have to take some tests to determine whether my heart is acting up or what kind of glasses I need. Similarly, I need to know if my algebra student is grasping the concept of solving equations, and the best way to do that is to test her, either formally on a written test, or informally as I watch her do a problem.

Since only God is all knowing, we mortals will never escape the need for tests, for testing in its most basic form is simply asking questions in order to know.

We may want to find out information for ourselves, or we may want to see how well someone else has processed information. As parents and teachers, we cannot know how well our students are doing unless we test them. The curriculum we use probably has ready-made tests designed to be used at various intervals in the course. We may design our own tests, depending on how comfortable we are with teaching. We may quiz our students verbally throughout the day. And then of course there are achievement tests, required in some states and not in others.

In this group of articles, we will look at tests in general, not just achievement tests. There is a lot of information on the web and in print, but it may be difficult to find, misleading, or loaded with eight-cylinder terminology. We want to keep these articles informative and understandable, and will try to explain terms as we go along.

TESTING TERMINOLOGY

There are many specialized terms relating to testing and item development. We are introducing some of them here, with an attempt to be accurate, but not intimidating. Terms are clustered by concept rather than alphabetically.

- **Attribute:** One thing we need to remember is that we do not measure or test a person or object. We are seeking to measure or evaluate the *attributes* of that person or object. Those attributes may be easily observable to our five senses and therefore easy to measure, or they may be much more abstract and correspondingly more difficult to measure.
- **Construct:** A specific skill or knowledge to be evaluated that has been formulated (*constructed*) and defined. A construct is somewhat abstract. As such, it is different from a concrete, obvious attribute of a person or object, such as height, weight, or speed, which can be measured objectively, accurately and with totally consistent results. We will use the terms *construct*, *concept*, and *learning objective* somewhat interchangeably.
- **Learning Objective (LO**): A specific goal of knowledge or skill. For example, "The student will be able to count successfully by tens from 10 to 100."
- **Construct Irrelevant Variance (CIV)** occurs when an item or test does not accurately measure the concepts (constructs) it is intended to measure because of errors in design or administration. For example, a student misinterprets an unclear item stem and selects the wrong response. CIV is one of the biggest threats to a test's validity.
- **Construct irrelevant mistakes, concept irrelevant errors, and similar synonymous phrases:** Errors students make while testing that are not directly related to the subject matter being tested. This includes misunderstanding of directions, carelessness, mislabeling of answers, improper marking of answer sheets, and other procedural errors.
- **Constructed-response / Selected-response:** Test item formats either require a student to come up with an answer or to select an answer from a list. Formats such as short-answer (fill-in-the-blank) and essay questions are known as *constructed-response* formats. Formats that require the student to choose the correct or best answer(s) are known as *selected-response* formats. Examples of these are multiple choice, true/false, and matching.
- **Item stem:** the part of an item that states the problem and asks for a response
- **Item response (or option):** The possible answers that a student must select from, typically the correct or best answer and one or more incorrect responses, known as distractors
- **Norm referenced:** A test that grades primarily by comparing a student's performance with others of his peer group. Scores are typically reported as percentile scores, grade equivalents, or stanines.
- **Criterion referenced:** A test that grades a student's performance by comparing it with specific defined criteria, usually a list of concepts the student is expected to master. Scores are typically reported as percentage correct, pass/fail, or letter grades.
- **Scoring Rubric**: A set of guidelines for scoring items – especially constructed-response items – clearly stating what to look for and how much credit to assign to each aspect of the answer. For example, a student may be assigned to come up with a pie chart for a set of data provided. A clear scoring rubric will define out of a total of 10 points for the item, how many points are awarded for correct percentages calculated, how many points are awarded for accurate piece sizes, how many points are awarded for clear, accurate, and complete labeling, how many points are awarded for neatness, and so on.

TEST ITEMS

An item is the basic unit of interaction on a test. What we often call test *questions* are more properly known as *items*, since a question is not always posed. The student's feedback is also more properly known as a *response* rather than an *answer*, but we won't get too particular on that point. Items can be written in various formats, including multiple choice, matching, true/false, short answer, and essay. We will only discuss the first three formats in this article, since they are more commonly used in achievement tests.

Since items are the actual points of interaction of students with the test, item quality is probably the most recognizable indicator of the overall quality of the test. High quality test items take time and effort to write, but are essential to a valid test. Items must test skills and knowledge of the subject at hand, not the student's test taking skills.

Seven Characteristics of a Good Test Item

1. **A good test item is relevant.** It should test the learning objective(s) being measured; nothing more and nothing less. This may sound obvious, but when a student who is highly skilled at taking tests scores better on an item than one who is less skilled, *even though he has no more knowledge on the subject*, this principle is probably being violated.

2. **A good test item is important.** Items must clearly address learning objectives, not trivia. Memorization of obscure facts is much less important than comprehension of the concepts being taught. Trivia, on the other hand, should not be confused with "core" knowledge that is the foundation of a successful education. Examples of "core", nontrivial knowledge include multiplication facts, common formulas, and common geographic names.

3. **A good test item is comprehensible.** Reading difficulty and choice of vocabulary should be as simple as possible relevant to the grade level being tested. This is a corollary of Characteristic #1. If you are not testing reading skills with an item, then do not make reading the item part of the problem. A good author is invisible; that is, you can read his story without being distracted by the style or skills of the storyteller. In the same way, the wording of a good test item should be "invisible". It should be simple, clear, and not a distraction from the concept at hand. In addition, because of this principle, there should be no objection to an item being read verbally to reading impaired students. This, of course, assumes that the item is not intended to evaluate reading skills.

4. **A good test item is unambiguous.** If a word has more than one possible definition, the context in which it is used should leave no reasonable doubt as to which definition is intended. Directions also should contain no ambiguity. If the student is to *circle* the correct answer, he should not be instructed to *mark* the correct answer.

5. **A good test item is straightforward.** There should be no trick questions. Tricky items often turn on the meaning of a single word that is not the focus of the item. This is often a flaw in true/false items. Use of the words *always* and *never*, and opinions stated as facts are often an

unneeded source of confusion to test-takers. If the correct response hinges on a single word, that word should be clearly emphasized. Humor should be used with care as well. The personality of an individual teacher may shine through in the tests he gives his students, but for serious or high-stakes tests, any attempt at humor can be confusing and distracting.

6. **A good test item is uncontroversial.** Items should be supportable facts or qualified opinions, not unqualified opinions. This principle is closely related to Characteristic #5. For selected-response items, there should be an unarguably correct answer. If more than one option could possibly be correct, the directions should call for the *best* answer, rather than the *correct* answer.

7. **A good test item is independent.** Items should not provide clues to the answers of other items. Sometimes a series of comprehension items all relate to a single reading passage, or multiple math problems are taken from a single scenario. This approach simplifies item-writing and can be effective, as long as the individual items are still independent of each other. On the other hand, if getting the correct answer on Item #2 depends on getting the correct answer on Item #1, then item #2 tells you absolutely nothing about the skills of the student who missed Item #1. Furthermore, this student is being penalized twice, in effect, for one mistake.

ITEM FORMATS & STANDARDS

Types of Test Item Formats

Test items can be written in various formats, including *multiple choice, matching, true/false, short answer,* and *essay*.These formats vary in their strengths and weaknesses, and no one format is ideal in all circumstances.

The first three formats are known as *selected-response* formats, because the student sees the possible answers and has to choose (or *select*) the correct one.

1. **Multiple Choice**. The student is given directions or a question, and needs to choose the correct or best answer from several possible responses.
2. **Matching.** The student is given two lists of words or phrases, and for each item in the first list, must choose the correct item from the second list to go with it. The matching format is actually a type of multiple choice, where each test item has the same list of possible responses.
3. **True/False.** This venerable format presents a statement, and the student must mark whether it is correct or not. It may take the form of true/false, yes/no, agree/disagree, or others.

The last two formats are known as *constructed-response* formats, because the student has to come up with the answers on his own.

1. **Short Answer.** The student is given a statement that requires him to fill in the blank(s), either in the statement itself, or at the end of it. As *short answer* implies, the expected response is usually a word or phrase. This format is common in curriculum-based tests.
2. **Essay.** This item format requires a longer answer that usually requires more creative thought or memory. The student may be asked to describe, discuss, or summarize a given subject. This format requires the most critical thinking skills and is also the most challenging for the examiner to score.

We will discuss only the three selected-response items for now.

The Multiple Choice Item Format

The multiple choice (MC) format is the most commonly used format in formal testing. It typically consists of a stem and three or more distractors, but can vary widely. The matching format can be thought of as an MC format, where several items share the same group of options. Multiple choice is popular for several reasons:

1. No subjective evaluation is required in scoring (the answer is either right or wrong, best or not best, not half-right or partly wrong).
2. It lends itself to detailed analysis of responses, in which even incorrect answers can provide information on the student's skills.
3. It lends itself well to computer scoring.

There is also one significant drawback to multiple choice. As a selected-response format, it is unable to test writing skills, including organization of thought and originality. Since these skills are generally beyond the scope of a standardized achievement test, it is not a serious issue in this context.

In addition to the general characteristics of a good test item noted above, there are some specific guidelines to follow when writing or evaluating MC items. Some relate to the stem, some to the options.

Characteristics of a Good Multiple Choice Item

1. The stem should clearly state the problem. A good stem is often clear enough that a knowledgeable student can answer the item correctly without seeing any of the options.
2. The stem should contain as much as the item as possible, but no more. There is no point in redundantly repeating something in each option that can be stated in the stem. On the other hand, the stem should not wordy nor contain irrelevant information, known as *window dressing*. One exception would be a problem presented that requires the student to determine which facts presented are necessary to solve the problem and which should be discarded.
3. The stem should, in most cases, be worded positively and in the active voice. When negatives do need to be used, they must be accentuated in **boldface** or ALLCAPS.
4. Use "story problems" – literally or figuratively – to present scenarios that require comprehension and analysis, not merely recall of the concept.
5. Always keep in mind that the primary goal in writing the response options in MC is to make it difficult for an uninformed person who is skilled at testing to figure out the correct answer. Knowledge of the construct being evaluated ideally plays the only factor in correctly answering an MC or any other item format.
6. Three or four options are best. It is difficult to write more than two or three plausible distractors. The various authors of the *Handbook of Test Development* range from mild to strongly-worded support of only three options.
7. All options should be parallel in structure and similar in length. The item is more readable, and there will be no obvious clues as to which options may be correct or are obviously incorrect.
8. Options must be grammatically consistent with the stem in order to prevent elimination of distractors.
9. All options must be plausible. If someone skilled, or at least comfortable, in a testing environment, were to take a test on a subject of which he knew nothing, he should not be able to dismiss options that seem to be implausible.
10. Distractors should reflect typical student errors, which makes them more plausible and more valuable in analyzing student performance.
11. The option, "*All of the above"*, is confusing and should generally be avoided. The option, "*None of the above"*, should only be used when there is one absolutely correct answer, as in spelling or math.
12. Options should avoid clang associations, where the correct answer contains a word or phrase from the stem that the distractors lack.
13. Options should be placed in a logical order, such as numerical, alphabetical, or response length. On the other hand, placement of the correct response should be random. Any discernable pattern of correct answers can invalidate a test.
14. Options should not overlap each other; one option should not be a partial version of another.

The Matching Item Format

As was mentioned earlier, the matching format can be considered a type of multiple choice. The matching format is common in curriculum-based tests. It is sometimes used to good advantage and sometimes very poorly done. Some of the strengths of the matching format are:

1. It is easy to construct. Since options are used for more than one item, not nearly as much effort needs to be put into constructing each individual item.
2. It is compact in size. An individual item usually takes only a fraction of the space occupied by one conventional MC item.
3. It is usually time efficient for the test taker. He only needs to analyze one set of options for multiple items, provided the matching group is competently designed.
4. It is very useful for working with groups of homogenous items, for example, matching states with their capitals.

There can also be some serious weaknesses in the matching item format, which could make an entire section of test items invalid. Some things to look out for:

1. Cued answers. A competent test-taker can usually get one or more items correct "for free", by using the process of elimination. A group of ten items with ten options often means that a student needs to know, at most, the answers to nine of the items.
2. Non- homogenous options. Many, many groups of matching items are practically worthless because they mix totally unrelated things together as options. In such cases, a skilled student can use the process of elimination to dramatically increase his score, and very little valid testing has taken place.
3. Excessively large groups of items or options. Since each item has the entire set of options as answer possibilities, a student may become overwhelmed with the amount of choices given from which to select the correct answer.

Recommendations for the Matching Format

1. There should be more options than items. This will reduce the effectiveness of elimination and guessing.
2. Even better, the group of items should be designed to use some options more than once and some options not at all. This will nullify the process of elimination completely. If this is done, it must be explicitly stated in the directions. For example: "Answers may be used once, more than once, or not at all."
3. If options may be used more than once, the pool of options can be much smaller and less confusing. Some very effective item groups of ten or more may have only three options for the entire group. An example would be listing various geographic and political characteristics of North American countries and having Canada, the United States, and Mexico as the three options.
4. Options must be homogenous. Do not mix crops with rivers or Roman numerals with geometric shapes. Note that *items* do not necessarily need to be homogenous, as long as the list of *possible answers* is. The idea is to prevent elimination based on test-taking skill.
5. Items and options should be ordered alphabetically or in some other logical arrangement. As in multiple choice, correct answers should form no discernable pattern.

The True/False Item Format

The true/false (T/F) format is limited in usefulness compared with most other formats, but is still common. A few reasons for its refusal to fade into oblivion are the relative ease of writing a true/false item and the ease and objectivity of scoring it. There are more problems than benefits, however:

1. T/F items tend to focus on trivial facts, rather than significant concepts. As a result, they tend to be either too easy or unreasonably difficult.
2. T/F items are much more likely to be ambiguous or "tricky" to answer. Often the answer turns on a single word. A student may need to analyze multiple words in the item to catch the one that is incorrect.
3. T/F items are too rewarding for guessers, since a random answer has a 50% chance of being correct. On a curriculum-based test, where a passing score typically is 75% - 80%, a chance of 50% may not be enough to boost the overall test grade. On a norm-referenced achievement test, guessing with a chance of 50% may significantly affect the overall score.

Suggestions for True/False Items

1. Avoid vague, indefinite, or broad terms in favor of precise statements. Good test items must be unambiguous, and T/F items even more so.
2. If the correctness of a statement hinges on a particular word or phrase, highlight or emphasize that word or phrase.
3. Avoid negative statements if at all possible. Negative statements are harder to decode, particularly those with two negatives.
4. Include similar numbers of true and false items and make them similar in length.
5. Group T/F items under a common statement, story, illustration, graph, or other material. This reduces the amount of ambiguity possible, since the items come from a specific frame of reference.
6. Avoid generalizations such as *all, always, never*, or *none*, since they usually trigger a false statement. Also avoid qualifiers like *sometimes, generally, often*, and *can be*, since they are often indicators of a true statement.

THE HUMAN FACTOR: TEST TAKING SKILLS & ERRORS

Presented in a Q & A Format

1. **Should students be penalized on a test when they make mistakes not directly relevant to the concept being tested?**
   Examples: circling an answer when writing it down is indicated, pressing down too lightly on an answer sheet, going outside the bubble, mislabeling a math answer (eg, ft instead of ft^2)

   Answer:
   a. It depends on the teacher's objectives.
      i. If part of the overall course objective includes the learning and application of good study and test-taking skills, what appears on the surface to be irrelevant may not actually be so.
      ii. If exercising good study and testing skills is a part of the course or class objectives, has the teacher made that clear to the class?

   b. It depends on the clarity of testing instructions.
      i. Are procedural instructions given clearly, unambiguously, and illustrated, if appropriate?
      ii. If the directions are written, is there extra attention given to reading-impaired students to make sure they understand?
      iii. Is the penalty for common mistakes, such as mislabeling math problems, clearly stated?

   c. It depends on the stakes of the testing
      i. If stakes are high, such as passing or failing a grade level based on test outcome, concept-irrelevant mistakes should be penalized lightly, if at all. Are you going to retain a student because he did not read and/or follow directions, or because he is struggling with concepts he needs to successfully master?
      ii. If stakes are fairly low, such as a lower grade on a report card, procedural errors may be penalized more strongly to reflect the less-formal objectives of the class as well as the formal objectives.

2. **Should students be taught test-taking skills?**

   Answer:
   Some test-taking skills are essential, whether a test is well-designed or not. Students must learn to read or listen to directions carefully, and to follow them precisely. Any skill that helps students to avoid construct-irrelevant mistakes is worthwhile. Examples include:
   a. Removing mental and physical distractions
   b. Approaching the test with a positive attitude
   c. Reading directions carefully
   d. Asking for clarification if any directions seem unclear
   e. Checking back over answers for mistakes
   f. Passing over difficult items and coming back to them later, if time permits

3. **What about skills designed to take advantage of the weaknesses of poorly designed tests?** These would include looking for:
    a. Clues from other test items
    b. Options that are obviously different from the other alternatives
    c. Eliminating grammatically inconsistent options
    d. Using elimination as cues for matching and multiple choice items

    If a test exhibits flaws that a good test-taker can exploit, the responsibility lies with the test, not the student. Any such skills can hardly be considered unethical.

4. **What about guessing skills?** These would include:
    a. Eliminating distractors determined to be incorrect
    b. Using partial knowledge to identify possible correct answers
    c. Relying on hunches and first impulse responses
    d. Randomly selecting answers if out of time or with a complete lack of knowledge

    Answer:
    Guessing skills are completely ethical, in my opinion. A good test will take guessing into account, and try to minimize its effect. Guessing can, however, reflect partial skill or knowledge, which is directly relevant to the concept being tested.
    Additionally, I do not believe that a good test will forbid guessing in any form. No struggles of conscience should add to the stress of taking the actual test.
    Randomly marking answers is usually discouraged, as test designers intend students at least to read the problem. It may be difficult to identify, however, unless the student has taken to marking interesting patterns on his answer sheet.

5. **What about cheating on a test?**

    Answer:
    Cheating is not a skill; it is a moral failure that undermines the validity of the test. Cheating is also a concept-irrelevant error, since it produces grades that do not reflect the student's actual knowledge. Cheating is not the fault of the test designer, although there are ways to combat cheating in test design and administration.

    Cheating can take several forms:
    a. Obtaining prior unauthorized knowledge of test content
    b. Copying answers from another test-taker
    c. Bringing "cheat sheets" in various forms to the test
    d. Unauthorized review of study materials during the test
    e. Exceeding time limits
    f. Using calculators when not permitted
    g. Using phones or other devices to get unauthorized help from others

    Cheating skills should obviously not be taught nor tolerated in a Christian environment or, for that matter, in any other environment.

TESTING READING-IMPAIRED STUDENTS

Students with reading difficulties usually do not score well on achievement tests. We have had many questions posed over the years about grade placement for such students and if it is even worthwhile testing them. Here are several points to consider:

Students with reading difficulties are often not mentally handicapped. They may not even have an overall academic learning disability. I have seen students who have struggled with reading in school who have gone on to excel in business and other environments. There are children who simply struggle with decoding words. These students tend to have a much better comprehension of what is read to them than what they read themselves. Often for students like that, a traditionally administered achievement test yields little information of value.

Some achievement test sections do not directly measure reading skills, such as math and language. Even some reading sections test mental understanding of a reading passage more than the actual decoding of words. Since the purpose of an achievement test is to accurately assess a student's skills, it is consistent with good testing practice to eliminate any factors that hinder such a valid assessment. In fact, the *Standards for Educational and Psychological Testing* devotes an entire chapter to the rationales and guidelines for testing individuals with disabilities. The overarching principle is that extraneous factors and disabilities should not affect the assessment of the actual learning objective to be measured (Standard 10.1).

We recommend some experimentation with struggling students. One possibility is having an assistant read test items to the student, but avoid any help with actual answer selection. Another is to extend time limits for the slow reader. Yet another is to test an easily-distracted student in isolation. Because of the nonstandard method of test administration, test results obtained by these methods should probably not be included in group scores.

If you are really determined to find out how much of a student's academic difficulties are due to reading problems, administer the test twice, using both Form A and Form B of the same level. The same person should administer the test both times. All other factors, such as time, day(s) of the week, and testing environment should be duplicated as much as possible from the first testing session to the second. The first time through, use verbal assistance extensively, including reading test items to the student. The second time, use the other form of the test and again supervise closely, but do not read actual test items to the student. The difference in percentile scores on the two tests should give some indication as to how much a reading disability affects the student's performance in other subject areas.

Questions often arise about grade placement for reading- or learning-disabled students. There is no point in testing a 6th grader who is working at a 3rd grade level by handing him 6th grade testing materials. Test a student at the grade level in which he can reasonably be expected to perform.

It is true that there are workarounds in a test, such as testing 3rd graders at the end of the year as beginning 4th graders. What does not work, however, is trying to go outside the boundaries established by the test. We have had students identified as 6th graders who have taken a Level 2 (grades 2-4) test. There are no provisions and no norms in the 1970 CAT for testing out-of-level grades. Therefore, such a student needs to take a Level 3 (grades 4-6) test. If Level 3 is

deemed too difficult to be meaningful for that student, he would need to take the level 2 test as a 4th grader.
One more note for our customers: When returning tests taken by learning-disabled students, a grade level must be provided for them, one that is appropriate for the test level they are taking. We cannot enter tests into our system without a grade level, since percentile scores and other norms are dependent on them.

We have tried to stress over the years that achievement tests are not some fearful, omniscient entity that must be approached with fear and trembling. Rather, they are carefully-developed tools that help schools, teachers, and parents to make informed decisions about the education of their children. Like any other tools we use, we are responsible to learn about their use and misuse, what they are intended to do, and what they cannot do. Like many tools, they can sometimes be used in more original ways than we first thought, provided that we are thoroughly familiar with them.