How to Prepare Better Tests:

Guidelines for University Faculty

Beverly B. Zimmerman
Richard R. Sudweeks
Monte F. Shelley
Bud Wood

Brigham Young University Testing Services
and
The Department for Instructional Science

Introduction

How good are your tests? As a teacher, you assume that test scores are based only on student ability and that they provide accurate information about student performance. Research has shown, however, that under certain conditions, these assumptions are wrong.

The adage, "if it ain't broke, don't fix it" obviously applies to testing. However, if you are a typical college professor, you may not be aware of some of the pitfalls that regularly occur in testing. This booklet will help you determine if your tests need improvement and will provide suggestions for making your tests more effective.

How can I use this booklet?

This booklet answers key questions that faculty members often ask about how to improve their tests. You can read the booklet in its entirety, browse through it and read only those specific questions that interest you, or use the summary checklist as a self-evaluation. The booklet also provides a list of suggested resources for additional information on testing.

Some suggestions in this booklet apply specifically to those tests administered and / or scored by the Testing Center. These suggestions are boxed.

TABLE OF CONTENTS

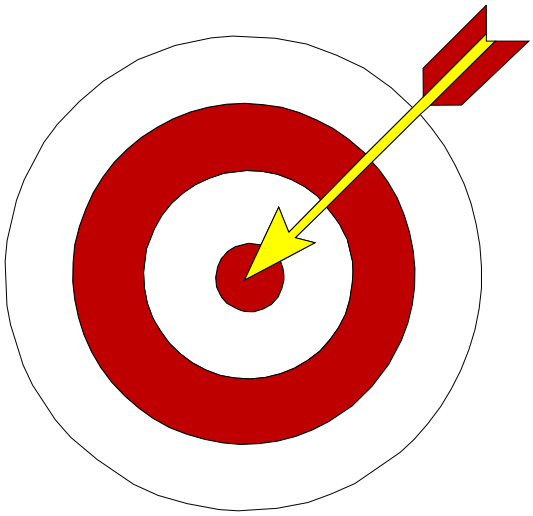
I. Developing a Test

How can I develop a good test?

Most classroom tests are developed for one or more of the following purposes:

- To establish a basis for assigning grades.
- To determine how well each student has achieved the course objectives.
- To diagnose student problems for remediation.
- To determine where instruction needs improvement.

Developing a good test is like target shooting. Hitting the target requires planning; you must choose a target, select an appropriate arrow, and take careful aim. Developing a good test also requires planning: you must determine the purpose for the test, and carefully write appropriate test items to achieve that purpose.



What are some of the pitfalls in developing a test?

One of the biggest pitfalls is that the target you hit may not be the target for which you aimed; that is, the test scores may not measure what you want them to. For example, if all of the test items require students to merely regurgitate factual information, the test will not likely provide as valid a measure of students' ability to apply their knowledge or to solve problems.

What can I do to increase the likelihood of obtaining more valid test scores?

You should make sure your instructional objectives and the test items are congruent. For example, if one of your instructional objectives is for students to apply principles to new

situations, then your test items should match your objectives by providing opportunities for students to exhibit that behavior.

One simple method for planning a test consists of creating a table with the task across the top (i.e., the learning outcomes or behaviors a student should exhibit as a result of your teaching, such as knowledge, understanding, and application, etc.) and the course content along the side.

Each cell in the table corresponds to a particular task and subject content. By specifying the number of test items you want for each cell, you can determine how much emphasis to give each task and each content area.

CONTENT	TASK			TOTALS
	Knows Specific Facts	**Understands Concepts**	**Applies Principles**	
Newton's Laws of Motion	4	4	12	**20**
Types of Forces	4	2	7	**13**
Buoyancy	2	4	4	**10**
Acceleration of Gravity	2	3	5	**10**
Friction	2	2	3	**7**
TOTALS	**14**	**15**	**31**	**60**

A test, no matter how extensive, contains only a sample of the many possible test items that could be included. Thus, sample items should be as representative as possible of the various areas your testing.

When should I use essay tests and when should I use objective tests?

There are no absolute rules for when to use an essay test that allows students to select, organize, and supply the answer in essay form or an objective test that allows students to select the correct answer from a number of alternatives (as in matching and multiple-choice questions). Both the essay and the objective test can measure instructional objectives in specific content areas and both have advantages as well as limitations.

The Testing Center provides test administration services for both types of tests. Approximately 550,000 tests with at least one or more essay questions are administered annually in the Testing Center and are returned to the course instructor for scoring. The essay scores are then loaded back into Testing Center computer files to be merged with objective scores (if applicable) or to provide cumulative record-keeping services.

The comparative advantages of each type of test are shown below:

	Essay Test	Objective Test
Instructional objectives measured	Does not measure recall or knowledge of facts efficiently. Can measure understanding, application, and other more complex outcomes.	May be designed to measure understanding, application, and other more complex outcomes as well as recall.
Item preparation	Fewer test items; may require less extensive preparation.	Requires a relatively large number of test items; necessitates extensive preparation.
Sampling of course content	Generally, quite limited because a small number of questions.	Large number of questions permits a broader sampling of course content.
Structure of task	Less structured, but may be influenced by writing ability or by bluffing.	Highly structured, but may be subject to guessing.
Encouragement to Students	Encourages organization, integration, and effective expression of ideas.	Encourages development of broad background of knowledge and abilities.
Scoring	Time-consuming; requires use of special measures for consistent results.	Easily accomplished, with consistent results; usually marked only right or wrong.

What general guidelines should I follow when writing test items?

Effective test items match the desired instructional outcome as directly as possible. Follow these general guidelines to avoid problems in writing test items:

- Present a single clearly-defined problem that is based on a significant concept rather than on trivial or esoteric ideas.
- Determine an appropriate difficulty level for students.
- Use simple, precise, and unambiguous wording.
- Exclude extraneous or irrelevant information.
- Refrain from providing unnecessary clues to the correct answer. For example, test-wise students have learned that the correct answer generally: (1) is longer, (2) is qualified or is more general, (3) uses familiar phraseology, (4) is a grammatically perfect extension of the question itself, (5) is one of the two similar statements, or (6) is one of the two opposite statements. They have also learned that incorrect answers often: (1) are the first or last option, (2) are logical extremes, (3) contain language or technical terms unexpected, (4) contain extreme words like "nonsense" or "foolhardy," (5) contain flippant remarks, or (6) contain completely unreasonable statements.
- Eliminate any systematic pattern for answers that would allow students to guess answers correctly.
- Guard against cultural, racial, ethnic, and sexual bias. Items should not require presupposed knowledge which favors one group over another. For example, a test item that refers to a "fly ball" assumes knowledge of baseball and may favor males.
- Avoid test items that assume stereotyped behavior (such as portraying minorities in a subservient role).
- Use answers from open-ended questions given in previous exams to provide realistic distracters.

II. Preparing, Assembling, and Administering a Test

How can I prepare and assemble a test more effectively?

Careful planning will help you prepare and assemble your test more effectively. Following these guidelines will allow students to take the test with minimum difficulty.

- Provide general directions for the test. Include the amount of time allowed for the test, how the items will be scored, and how to record answers. Set off the directions by appropriate spacing or different type style.

- Arrange items systematically. If the test contains several types of items, group similar items (such as all multiple choice items) together. Provide a clear set of directions for each new group of items.

- Place the most difficult questions near the end of the test so that students have time to answer more questions.

- Provide ample margins. Cramming too many test items into a page will only result in inefficiency during the administration and scoring of the test.

- Number the items consecutively.

- Don't split the item onto two pages. Keep introductory materials and the space for answering on the same page.

- Place the alternatives to multiple-choice items in a vertical column beneath the stem of the item, rather than across the page.

- Number each page consecutively and indicate the total number of pages in the test. This prevents problems later when pages may become separated. Include a header such as the following <u>on each</u> page of the test:

  ElEd 354 John Q. Prof Page 2 of 4

- Make sure all copies of the test are legible and free of typographical or grammatical errors.

- Before administering the test, prepare answer keys and scoring procedures.

If the test is to be administered or scored by Testing Services, include a cover page containing the teacher's name and section number, and the Testing Center's bar code (located in the upper right corner of the test). For example:

ElEd 354

Test 3-B

Total Pages: 5 John Q. Prof

DO NOT WRITE ON THIS EXAM!

How can I arrange the test items to facilitate scoring?

The following guidelines for arranging test items will facilitate scoring in an efficient manner:

- Space test items so that they can be read, answered, and scored with the least amount of difficulty. Double-space between items.
- Place answer spaces for objective items in vertical columns for easy scoring with each answer space clearly associated with the corresponding item.
- Provide adequate space for students to supply short-answer questions. Provide a full page for answering lengthy essay questions.

What should I do to facilitate students' use of the test center?

In addition to the guidelines given above for preparing and assembling a test, you should do the following:

- Be sure you have scheduled your test for administration in the Testing Center before you inform your students it will be administered there.
- Deliver the test to the Testing Center by noon the day before you plan to give the test. This will guarantee the test is ready to be administered.
- Make sure you have included enough copies of the exam.
- Schedule the test early, provide plenty of days for the exam, and try to avoid scheduling the exam on days when the Testing Center is crowded.
- Give students an indication of how much time the test will take so that they can plan their time accordingly.
- Notify students of any test late fee, telling them the days the fee will be charged and how much the fee will be.
- Inform students of the Testing Center web pages available at http://testing.byu.edu or through Route Y. Encourage them to visit these pages to verify Testing Center hours and to avoid long lines.
- Notify students if they must use the calculators supplied by the Testing Center, so they won't be surprised to learn they aren't allowed to use their own calculators.
- Inform students concerning the materials (such as textbooks, dictionary, calculator, etc.) they are permitted to bring in or are prohibited from bringing into the Testing Center.

What can I do to keep scores from being influenced by factors other than student ability?

It is important during the administration of the test to control those factors (other than student ability) that would influence test scores. You can lessen the chance of scores being influenced by outside factors by doing the following:

- Maintain test security. Make sure students cannot obtain copies of the test before it is administered. Keep all tests in secure locations.

- Take measures to reduce cheating by asking students to space themselves with an empty desk in between students on the same row, if possible; being sure the test administration is supervised at all times; and by explaining in advance the consequences for cheating if it occurs.

Testing Services supervises all tests administered in its facilities. Testing Center policy regarding cheating is to dismiss from the Testing Center any student caught cheating and to inform the teacher and the Honor Code office in writing of any irregularities.

- Don't provide unfair help to individual students who ask questions during the exam. Only provide information relating to the administration of the exam, not to the content of the test items.

- Provide sufficient time for students to complete the test. It more than 10% of the students fail to complete the test in the allotted time, time has become a factor influencing the test scores.

III. Evaluating a Test

How can I appraise the effectiveness of my test?

You do not have unlimited time and money to develop the "perfect" test for your course and purpose; however, you can still evaluate your test to determine if it's the best test you could develop, given limited time and resources. Like most written work, the first draft of the test will probably need to be edited and revised.

Here are some things you can do (both before the administration of the test and after) to appraise the effectiveness of the test.

Before administering the test:

- Obtain feedback about the effectiveness of your test from other sources. For example, colleagues can provide valuable input on the importance or appropriate difficulty of items; editors can evaluate the clarity of the test; testing specialists can determine factors in the test format that might raise or lower student scores; and students can indicate whether the material was covered in class.

- Use the checklist provided in this booklet to evaluate your test.

- Create a table showing the number of items for each content area and each task (similar to the table on page 3 of this booklet) to determine if all areas of content have been tested adequately.

- Give a small random sample of the test to students.

- Add non-graded experimental items to the test. The items can be used to pretest specific items for a final examination or to obtain feedback from students regarding the current test.

- Revise your test as time allows.

After administering the test:

- Review statistical feedback from the Testing Center about the effectiveness of the test.

- Get student input during class discussion of the test. Write down suggestions for making the test more effective, then revise items for future use.

- Consult other sources–published tests booklets, the references listed in this booklet, and the workshops and seminars sponsored by BYU Testing Services–for additional ideas and information and how to improve test items.

What statistical feedback can BYU Testing Services provide?

If your test is administered or scored by BYU Testing Services, you can receive an analysis of the following questions:

- How did the class as a whole score on the test? The analysis plots the scores from highest to lowest and indicates the number of students taking the test and the number of students receiving each score.
- How did the typical or average student score on the test? The analysis calculates the class mean (average score). A low class average may indicate the test was too difficult, while a high class average may indicate the test was too easy.
- How much variability (or spread) is there in the scores? The analysis provides the standard deviation and the interquartile range. The larger the value of each of these statistics, the greater the dispersion of scores.
- How reliable is the test? The analysis calculates an estimate of the consistency with which you would get comparable results if the test were given repeatedly or if two items representing the same content were given one after another.

What is an acceptable reliability estimate?

A higher reliability estimate indicates that the test results would be the same if administered again to the same (or similar) group of students. If you're using the test scores to determine what to review, or if you'll be testing the student again, a lower reliability may be acceptable. Generally, however, a reliability estimate less than 0.60 indicates that the test may have problems with consistency and would not to produce comparable results if given again.

What can I do to increase the likelihood of obtaining higher reliability estimates?

You can increase the likelihood of obtaining higher reliability estimates by doing the following:

- Increase the length of tests. Longer tests measuring the same tasks and content–provided the test items are of high-quality–are more reliable.
- Write test items that measure more complex learning outcomes.
- Test your students more often. Shorter intervals between tests usually prevent large fluctuations in students' scores.

- Grade the test as objectively as possible. Make sure all test items (including essay questions) are scored accurately and without bias.

- Keep test at an appropriate difficulty level.

How can I determine when a test item needs to the revised?

The only practical way is to analyze each item after the test is given. If the Testing Center administers or scores your test, you can receive an item analysis indicating the following: (1) the percent of the total number of students who got the item correct (item difficulty), (2) how many more high-scoring than low-scoring students got the item correct (the item discriminating power), and (3) how many students answered each response to the item.

The following guidelines should help you determine when a test item should be revised:

- Items that more than 90% of the students answered correctly may be too easy. Low item difficulty may suggest that the item contained clues to the answer.

- A low discrimination grade for an item suggests that poor students scored higher than the better students on this question. This can be caused by an incorrect answer key or a poorly worded question. If the problem is an incorrect answer key, Testing Services can re-grade the test.

- A low percentage of students answering an item may indicate students ran out of time during the test or they were unsure of the answer and didn't want to guess.

- Items that fewer than 30% of the students answered correctly may be too difficult. A large number of difficult items may suggest the need for further instruction or for revision of the test.

IV. Scoring a Test

What problems can occur in scoring a test?

Scoring procedures must be accurate and yield uniform results regardless of the scorer; however, any scoring procedure is subject to error. For example, hand scoring can result in errors in counting answers or assigning points. (If you hand-score your tests, contact Testing Services for suggestions to minimize errors.) Machine scoring can be affected by students making improper changes in answers or by stray marks on the form. Essays scoring can be influenced by the fatigue level of the grader, by the order in which the papers are graded, or by the student efforts at "bluffing."

No matter what the scoring procedure, it is important that you use a grading key or establish a standard set of rules for assigning points for each question.

How can I design a scoring method for essay questions that will maintain objectivity and fairness?

The following guidelines will help yield accurate results while scoring essay questions:

- Cover the student's name while scoring the exam.
- If someone else does the scoring, provide supervised practice in scoring on sample papers.
- Have each paper re-scored by a second person. If this is impractical, re-score a sample of the papers (such as every 5th paper) to determine if additional re-scoring is necessary.
- Score all the papers for one essay question as a group before going on to the next question and cover the score so that it can't be seen when you're scoring the next question. Research has shown that scorers can be influenced by how students performed on an earlier question.
- Randomly rearrange all the papers after grading all the answers for an individual question so that scorers are not influenced by the order of arrangement of the papers or by fatigue.
- Determine before hand how to score papers with extraneous information or irrelevant errors, such as punctuation or grammar errors.
- Determine how to record test scores.

The Testing Center can provide you with a disk (for use with special software available from the Testing Center) to facilitate the loading of essay scores to cumulative record files. Scores loaded in this manner can then be weighted and merged into students permanent records.

How can I design an effective scoring method for objective questions?

Before you score objective questions, you should determine the following:

- Whether there is one best answer, or multiple correct responses.
- How many points students will receive for each item. Normally, each item correct receives one point, but important questions might receive more points.

> If you indicate a zero ("0") weight for an item, Testing Services will not score that item.

- Whether you will penalize guessing.
- Whether to group scores into subgroups to provide information to students about how they scored in specific content areas.

V. Interpreting and Using Test Results

How are test scores commonly misinterpreted?

Teachers may misinterpret test scores by assuming that high scores indicate good instruction and low scores indicate poor students. Students may misinterpret test scores by assuming their high scores indicate they are smart or well prepared and their low scores indicate poor teaching or an inadequate textbook.

High scores can result from a test that is too easy, from only measuring simple instructional objectives, from biased scoring procedures, or from other factors that influenced the scores, such as cheating or providing unintentional clues to the right answers. Low scores can result from a test that is too difficult, from trick questions, from testing content not covered in class, or from other factors that influence the scores, such as grader bias or insufficient time to complete the test.

What can I do to lessen the likelihood of misinterpreting test scores?

Interpreting test scores can be difficult. The following guidelines will help lessen the likelihood of misinterpreting test scores:

- Avoid interpreting small differences in students' scores as being indicative of real differences in their level of achievement.

- Assign final grades based on more than 1 or 2 exams. Include their performance on projects, lab assignments, research papers or proposals, and other written work, products, or ideas.

- Avoid comparing scores on two different exams unless the exams have been mathematically equated. Even multiple forms of the same exam may differ significantly.

Testing Services can mathematically equate tests for you.

- Refrain from using grading practices that strictly adhere to 90% = A, 80% = B, etc., regardless of the difficulty of the test.

- Evaluate the tests using the guidelines provided in the previous section of this booklet.

How can I use the test results to improve learning?

The following are several ways to use test results to improve learning:

- Return the corrected tests to the students in a timely manner and discuss the test with your students.

- Give regular quizzes to help students keep up on the reading.
- Allow retakes of earlier tests (using a parallel form of the test) to encourage students to improve their understanding before going on to new concepts.
- Look at subtest scores and item analysis information to identify areas where the course may be weak or where students may need help.
- Grade essay exams with two scores–one for content and one for grammar–to encourage students to improve writing skills.
- Provide a short questionnaire at the end of the test to allow students to provide feedback about the course and the exam.
- Have each test include review items from previous units.

Where can I obtain additional help?

You can obtain additional help by referring to the books listed in the bibliography of this booklet, and by attending seminars and workshops sponsored by Testing Services. Contact Testing Services to indicate topics of interest to you and to receive notification of upcoming workshops. Testing Services can also provide qualified graduate students to help you evaluate the tests, or assistance on research questions requiring testing.

What should I do now?

You may want to review the principles discussed in this booklet by using the following summary checklist each time you prepare a test.

Checklist for Effective Tests

The Test Development

- ❑ Do the test items match the course objectives?
- ❑ Do the test items relate to what was actually taught?
- ❑ Do the test items measure important concepts rather than trivia?
- ❑ Do the test items measure more complex behavior, such as understanding of basic principles or ability to make practical applications, rather than simply measuring recall?
- ❑ Are the test items free from vaguely defined problems, ambiguous wording, extraneous or irrelevant information, and unintentional clues to the correct answers?

Test Preparation and Administration

- ❑ Do the test directions specify how the items will be scored and how the students should record answers?
- ❑ Are items presented in the same format grouped together?
- ❑ Are the items and pages numbered consecutively?
- ❑ Have I provided sufficient time for students to complete the test?
- ❑ Have I made provisions to reduce cheating?

Test Evaluation

- ❑ Was the test long enough to provide a valid, reliable estimate of the students' achievement?
- ❑ Are there means for grading students other than on the basis of this exam?
- ❑ If the purpose of the test was to rank students (rather than to assess mastery) did I reject items which nearly every student either missed or got correct?
- ❑ If the test were given again, can I feel confident the results would be consistent with current test scores?
- ❑ Have I considered student input regarding ambiguity and problems with specific test items?

Test Grading and Scoring

- ❑ Are the items spaced so they can be read, answered, and scored efficiently?
- ❑ Is each answer space clearly associated with its corresponding item?
- ❑ Have I established a grading key for all items, even essay questions?
- ❑ Have I made provisions to maintain student anonymity and to prevent grader bias?
- ❑ Have I checked for accuracy by providing two scorings or by re-scoring sample papers?

Bibliography

Bloom, B.S. (Ed.) (1956). *Taxonomy of educational objectives. Handbook I: Cognitive domain.* New York: David McKay Co.

Bloom, B.S., Madaus, G.F. & Hastings, J.T. (1981) *Evaluation to improve learning.* New York: McGraw-Hill, 1981.

Cashin, W.E. (1987). Improving essay tests. *Idea Paper No. 17.* Center for Evaluation and Development. Kansas State University. (ERIC Document Reproduction Service No. ED 298 151).

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N.E. (1988). *How to construct achievement tests* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N.E. & Linn, R.L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Hopkins, K.D., Stanley, J.C., & Hopkins, B.R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Making the classroom test: A guide for teachers. (1961) Princeton, New Jersey: Educational Testing Service.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn, (Ed.), *Educational Measurement* (3rd ed.), (pp. 335-366). New York: Macmillan.

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R.L. Linn, (Ed.), *Educational Measurement* (3rd ed.), (pp. 447-474). New York: Macmillan.

Sparks, W. G. (1987). The art of writing a test. *The Clearing House.* Vol 61 (December), pg. 175-178.