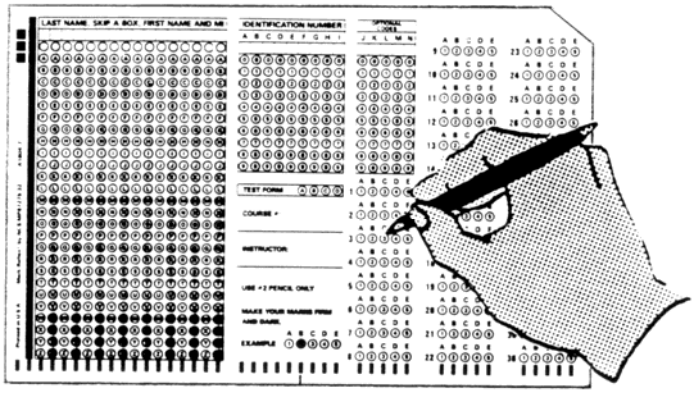
How to Judge the Quality of an Objective Classroom Test

Technical Bulletin #6



Evaluation and Examination Service
The University of Iowa
(319) 335-0356

HOW TO JUDGE THE QUALITY OF AN OBJECTIVE CLASSROOM TEST

The purpose of this bulletin is to acquaint instructors with the characteristics of a "good" objective test and to suggest procedures that may improve their tests. These guidelines apply most appropriately to tests that are designed to identify differences in achievement levels between students (norm-referenced tests). Some of the criteria outlined either do not apply or apply in somewhat different ways to tests designed to measure mastery of content.

Some of the important factors to consider in judging the quality of a test are indicated by these questions:

1. Do the test questions adequately reflect the course objectives?
2. Is the test fair to the students in view of the instruction given them?
3. Is the test administered under conditions that give all students an optimal and equal chance to demonstrate their achievement?
4. Does the test emphasize important, long-run achievements more than incidental, quickly forgotten information?
5. Is the length of the test appropriate for the time available--long enough to give reliable scores but short enough so most students have time to attempt all items?
6. Are the questions individually effective in distinguishing between high and low achieving students?
7. Is the test of appropriate difficulty, neither too hard nor too easy?
8. Does the test as a whole distinguish clearly between students at different levels of ability?
9. Are the scores reasonably reliable, so that they would agree closely with those from another equivalent test?
10. Are the scores reasonably accurate, so that they closely approximate the students' hypothetical true scores?

Answers to the first five questions depend largely on the knowledge and judgment of the instructor. The remaining five can be answered with the aid of a test analysis report, available to instructors through the Evaluation and Examination Service. Although the analysis is designed primarily for tests that have been administered to larger classes (30 or more students), instructors

in smaller classes may find the information helpful. Further references to the test analysis report will be made in the pages that follow.

TEST CHARACTERISTICS TO EVALUATE

MEASURES COURSE OBJECTIVES

Most courses are expected to make some permanent changes in the students who take them--to leave them with new knowledge and understanding; improved and extended abilities; or new attitudes, ideals, and interests. Often such goals are neglected when tests are constructed. Instead of being evaluated on the basis of ultimate objectives, students are often judged on the basis of what they remember or what they read in preparation for the test.

When several instructors teach the same course, it is common for them to share the responsibility for test construction. This practice could easily be extended to most courses in a department. Instructors in the same department ought to exchange examinations for review and constructive criticism. While it is true that good instructors present courses that are unique products of their own special abilities, the important achievements they teach and test ought to be things that most of their colleagues also would accept as true and important. Instructors should not feel obligated to accept and apply all the suggestions made by their colleagues, since the ultimate responsibility for the quality of the test is their own. However, independent reviews of a test by competent colleagues cost little additional time and can yield large returns in improved quality.

To ensure relevance of test content, a test plan usually is developed to guide the preparation of the test. The content to be covered and the relative emphasis to be given to each aspect of it are indicated in the test specifications.

FAIRNESS TO STUDENTS

A test is fair to students if it emphasizes the knowledge, understanding, and abilities that were emphasized in the actual teaching of the course. Additionally, a test's proportional emphasis on various aspects of the course ought to approximate their relative importance as previously conveyed to students through class time allocation, reading outlines, and lists of course objectives.

Instructors are in the best position to judge the fairness of a test given to their own students. Probably no effective test has ever been given that was regarded as perfectly fair by all persons taking it. On the other hand, student comments on fairness are often worthwhile for the instructor to obtain and to contemplate. A request for comments can show the instructor's concern for fairness. A student may call attention to ambiguity in a question, to the presence of questions that deal with matters not covered in class, or to the omission of questions on matters that were stressed. There are few classroom tests so good that they cannot be improved by attention to student comments and suggestions.

CONDITIONS OF ADMINISTRATION

Was the test handled efficiently without confusion or disturbance that could interfere with effective performance? Were all examinees on an equal footing as far as prior knowledge of the nature of the examination? Did they have enough prior knowledge to be able to prepare properly for it? Was cheating prevented? Were physical conditions of light, heat, and freedom of movement satisfactory? These questions are best answered by the instructor who gave the test. But here again, if any doubt exists about conditions of administration, student comments can be helpful.

IMPORTANT ACHIEVEMENT

Did the test emphasize important long-run achievements? Sometimes a test consists mainly of questions requiring recall of some detail in the process of instruction --"How did the lecturer illustrate Hooke's Law?"; or requiring reproduction of some unique organization of subject matter "What were the three chief reasons for the failure of the League of Nations?". Such items do not measure important achievements. If a majority of the questions deal with applications, understanding, and generalizations; if knowledge of terms and isolated facts is not the sole aim of a large proportion of the questions; and if questions deal with matters of value outside the classroom the test does emphasize important achievements.

TIME LIMITS

Most tests of achievement at the college level should be work-limit tests rather than time-limit tests. That is, students' scores should depend on how much they can do, not how fast they can do it. Speed may be important in repetitive, clerical-type operations, but it is ordinarily not important in critical or creative thinking or decision making. The fact that good students tend to be quicker than poor students is not a good reason for penalizing the occasional good but slow student. Hence it is recommended that test time limits be generous enough for at least 90% of the students to attempt and complete all questions in the test.

Of course a test that is so short that everyone can easily finish it may not be very efficient and may not yield very reliable scores. The larger the number of independent observations we can obtain in the testing time available, the more accurately we can measure the amount of achievement students possess. Within reasonable limits, a longer test will yield more accurate and stable scores than a shorter one.

ITEM DISCRIMINATING POWER

The discriminating power of an item can be represented by any of several indices of discrimination. A good discriminator is an item that high achieving students answer correctly and low achieving students answer incorrectly. One index used to represent this relationship is the difference between the proportion of good and poor students who respond correctly. For statistical reasons, those students in the top 27% in terms of total test score are taken to be good students, and those in the bottom 27% are taken to be poor students. If the discrimination index is .30 or above, the item can be said to be an effective measure of the same achievement that is measured by the entire test. The index is negative when more students in the lower group than upper group answer the particular item correctly. The more items classified as highly or moderately discriminating, the more reliable are the test and the resulting grades. It should be noted that an item discrimination value is unique to a group of examinees. An item with satisfactory discrimination for one group may be unsatisfactory for another. Further discussion of discrimination indices can be found in Technical Bulletin # 17, "Reading Your Test Analysis."

TEST DIFFICULTY

The difficulty index of a test item is the proportion of a particular group that answered the item correctly. Multiple choice items for which the difficulty indices are about .50 to .70 are ideal in terms of difficulty. If almost all the students taking a test get an item correct (or incorrect) then the item is not very efficient. That is, items that are extremely difficult or easy provide too little information about student achievement in relation to the amount of testing time they require. For the test as a whole, the average score should be about midway between the expected chance score and the maximum possible score. Item difficulty, like discrimination, is associated with a group of students. Items that seem too difficult for students just beginning a course probably will seem too easy for those who have completed that course.

LEVELS OF ABILITY

For a test to distinguish clearly between students at different levels of ability it must yield scores of wide variability, as indicated by the standard deviation. The larger the standard deviation (for a fixed number of items), generally speaking, the better the test. A standard deviation equal to one-sixth of the range between the highest possible score and the expected chance score is generally considered an acceptable standard for groups of 100 or more examinees. For some good tests, the standard deviation is larger than one-fourth of the available range, and for poorer tests it may be less than one-tenth of the available range. If a test is too hard, too easy, or composed of too many poorly discriminating items, it will yield scores having a small standard deviation. The size of the standard deviation of a set of test scores has an important bearing on the reliability of the scores.

TEST RELIABILITY

The reliability coefficient included in a test analysis report represents the estimated correlation between the scores on the test and scores on another equivalent test, composed of different items, but designed to measure the same kind of achievement. The highest possible value is 1.00. A high reliability coefficient indicates that a student's score was not overly influenced by measurement errors. Many good commercial objective tests have reliability coefficients of .90 or more. This level is difficult to achieve consistently with homogeneous

class groups and with items that previously have not been administered, analyzed, and revised. But a reasonable goal for instructors to set is a reliability estimate of .80.

The reliability of a test is affected by how well the items discriminate between high and low achievers; how many items there are; how similar the items are with respect to the ability measured; and how much the students differ from one another in the ability measured. Thus, it is possible to get more reliable scores in one course than in another and with one group of students than with another. Most of the factors that influence reliability are under the instructor's control. If the coefficient is too low it can almost always be raised by improving the discrimination of the items used, or by adding more items.

ACCURACY OF SCORES

The accuracy of the scores is reflected by the standard error of measurement (SEM), a statistic computed using the standard deviation and the reliability coefficient. This index may be interpreted as follows. If the SEM is 2 score points, for example, one can say that about two-thirds of the scores reported were within 2 points of each student's "true" score. About one-sixth of the students received scores more than 2 points higher than they should have received. The remaining one-sixth received scores more than 2 points too low. Unfortunately one has no way of knowing which student should have received higher or lower scores or how much higher or lower each student's score should have been.

The size of the SEM can be misleading. Test A, with a SEM of 2, would rank students more accurately than Test B, with a SEM of 1, if the standard deviation of scores for Test A was more than twice as great as that for Test B. Hence the reliability coefficient, which reflects the ratio between SEM and standard deviation, is a better indication of useful score accuracy than the SEM itself. The SEM simply serves as an indication of how much chance error remains in the scores from even a good test.

FINAL COMMENT

Test analysis does not in itself improve a test. But the analysis data make the prospects for improvement much greater for the instructor who is willing to be guided by the data. Weak items identified in the analysis can be discarded or revised. The coverage of the test can be extended and balanced. If the causes of unsatisfactory performance of the test as a whole or of

any item is obvious, they can be addressed by the instructor. If the causes are not so obvious, assistance can be sought from the staff of the Evaluation and Examination Service. The aim of systematic test analysis procedures is to make it as convenient as possible for instructors to evaluate their tests, and to improve them where needed.

The checklist printed on the next page can be used to assess the quality of an achievement test. The test characteristics noted can be judged from data found in the test analysis report and from an inspection of a copy of the test. Instructors might use this list to conduct an "internal" self-analysis or they might request an "external" review from a staff member at the Exam Service.

A more complete explanation of the statistical analysis provided by the Exam Service for multiple-choice tests is available in Technical Bulletin 17, "Reading Your Test Analysis."

CHECKLIST FOR EVALUATING A CLASSROOM TEST

(Norm referenced)

	Satisfactory	Needs Improvement	Not Applicable
TEST PLAN			
1. A test plan was developed	_____	_____	_____
2. Plan included detailed description of the content to be measured	_____	_____	_____
3. Plan identified the relative emphasis (percentage) to be given to each area of content	_____	_____	_____
RELEVANCE			
4. Tested important content (avoided trivia)	_____	_____	_____
5. Only relevant skills were measured - not reading, hand-writing, neatness, etc.	_____	_____	_____
6. Item content related closely to course objectives	_____	_____	_____
BALANCE			
7. Test covered all content areas noted in the test plan	_____	_____	_____
8. Test followed relative emphases established in test plan	_____	_____	_____
EFFICIENCY			
9. Most students finished in time allotted	_____	_____	_____
10. Test contained an adequate number of items	_____	_____	_____
RELIABILITY			
11. The reliability estimate (KR20) was adequate	_____	_____	_____
12. Items discriminated well (based on discrimination indices)	_____	_____	_____
13. All item options functioned well (wrong answers were plausible)	_____	_____	_____
ADEQUACY OF THE TEST ITEMS			
14. Items presented clear and definite questions or tasks	_____	_____	_____
15. Items of each type followed the rules for constructing that item type	_____	_____	_____
16. Items were free of ambiguity	_____	_____	_____
17. Items included only familiar vocabulary	_____	_____	_____
18. Items were of appropriate difficulty	_____	_____	_____
TEST IS TECHNICALLY SOUND			
19. Test was free of typing errors	_____	_____	_____
20. Instructions were clear and complete	_____	_____	_____
21. Keyed response position varied appropriately	_____	_____	_____
22. Exam copy was legible - attractive	_____	_____	_____