Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 4, Number 10, November, 1995 ISSN=1531-7714

Basic Item Analysis for Multiple-Choice Tests.

Jerard Kehoe,
Virginia Polytechnic Institute and State University

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

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

WHAT MAKES AN ITEM PSYCHOMETRICALLY GOOD?

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

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

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

The proportion of students answering an item correctly also affects its discrimination power. This point may be summarized by saying that items answered correctly (or incorrectly) by a large proportion of examinees (more than 85%) have markedly reduced power to discriminate. On a good test, most items will be answered correctly by 30% to 80% of the examinees.

A general indicator of test quality is the reliability estimate usually reported on the test scoring/analysis printout. Referred to as KR-20 or Coefficient Alpha, it reflects the extent to which the test would yield the same ranking of examinees if readministered with no effect from the first administration, in other words, its accuracy or power of discrimination. Values of as low as .5 are satisfactory for short tests (10 - 15 items), though tests with over 50 items should yield KR-20 values of .8 or higher (1.0 is the maximum). In any event, important decisions concerning individual students should not be based on a single test score when the corresponding KR-20 is less than .8. Unsatisfactorily low KR-20s are usually due to an excess of very easy (or hard) items, poorly written items that do not discriminate, or violation of the precondition that the items test a unified body of content.

IMPROVING THE ABILITY OF ITEMS TO DISCRIMINATE

The statistics usually provided by a test scoring service provide the information needed to keep a record of each item with respect to its performance. One approach is simply to tape a copy of each item on a 5 x 7 card with the test content

area briefly described at the top. In addition, tape the corresponding line from the computer printout for that item each time it is used. Alternatively, item banking programs may provide for inclusion of the proportions marking each option and item discrimination coefficients along with each item's content.

A few basic rules for item development follow:

1. Items that correlate less than .15 with total test score should probably be restructured. One's best guess is that such items do not measure the same skill or ability as does the test on the whole or that they are confusing or misleading to examinees. Generally, a test is better (i.e., more reliable) the more homogeneous the items. Just how to restructure the item depends largely on careful thinking at this level. Begin by applying the rules of stem and option construction discussed in ERIC Digest TM 95-3 (ED 398 236). If there are any apparent violations, correct them on the 5x7 card or in the item bank. Otherwise, it's probably best to write a new item altogether after considering whether the content of the item is similar to the content objectives of the test.

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

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

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

SOME FURTHER ISSUES

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

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

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

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

The reader may question whether it is ethical to throw out poorly performing questions when some students may have answered them correctly based on their knowledge of course material. Our opinion is that this practice is completely justified. The purpose of testing is to determine each student's rank. Retaining psychometrically unsatisfactory questions is contrary to this goal and degrades the accuracy of the resulting ranking. This article was adapted with permission from "Testing Memo 5: Constructing Multiple-Choice Tests--Part II," Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

FURTHER READING

Airasian, P. (1994) *"Classroom Assessment*," Second Edition, NY" McGraw-Hill.

Brown, F. (1983), "*Principles of Educational and Psychological Testing*," Third edition, NY: Holt, Rinehart & Winston. Chapter 11.

Cangelosi, J. (1990) "*Designing Tests for Evaluating Student Achievement*." NY: Addison-Wesley.

Grunlund, N. (1993) "*How to make achievement tests and assessments*," 5th edition, MA: Allyn and Bacon.

Descriptors: *Item Analysis; Item Banks; Measurement Techniques; *Multiple Choice Tests; *Psychometrics; *Scoring; *Test Construction; Test Items

Citation: Kehoe, Jerard (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, 4(10). Available online: http://PAREonline.net/getvn.asp?v=4&n=10.