

SCOREPAK®: ITEM ANALYSIS

Office of Educational Assessment
University of Washington
430 Roosevelt Commons B – Box 354987
56 Mary Gates Hall – Box 352807
Seattle, WA 98195-5837

e-mail: scorepak@u.washington.edu
<http://www.washington.edu/oea/score1.htm>
voice: 206.543.9899 / fax: 206.543.3961
voice: 206.616.7750 / fax: 206.616.9934

Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. Item analysis is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. In addition, item analysis is valuable for increasing instructors' skills in test construction, and identifying specific areas of course content which need greater emphasis or clarity. Separate item analyses can be requested for each raw score¹ created during a given ScorePak® run.

A basic assumption made by ScorePak® is that the test under analysis is composed of items measuring a single subject area or underlying ability. The quality of the test as a whole is assessed by estimating its "internal consistency." The quality of individual items is assessed by comparing students' item responses to their total test scores.

The remainder of this bulletin describes the various statistics provided on a ScorePak® ITEM ANALYSIS report. This report has two parts. The first assesses the items which made up the exam. The second part shows statistics summarizing the performance of the test as a whole.

Item Statistics

Item statistics are used to assess the performance of individual test items on the assumption that the overall quality of a test derives from the quality of its items. The ScorePak® ITEM ANALYSIS report provides the following item information.

Item Number. This is the question number taken from the student answer sheet, and the ScorePak® Key Sheet. Up to 150 items can be scored on the Standard Answer Sheet (purple).

Mean and S.D. The mean is the "average" student response to an item. It is computed by adding up the number of points earned by all students for the item, and dividing that total by the number of students.

The standard deviation, or **S.D.**, is a measure of the dispersion of student scores on that item, that is, it indicates how "spread out" the responses were. The item standard deviation is most meaningful when comparing items which have more than one correct alternative and when scale scoring is used. For this reason it is not typically used to evaluate classroom tests.

Item Difficulty. For items with one correct alternative worth a single point, the item difficulty is simply the percentage of students who answer an item correctly. In this case, it is also equal to the item mean. The item difficulty index ranges from 0 to 100; the higher the value, the easier the question. When an alternative is worth other than a single point, or when there is more than one correct alternative per question, the item difficulty is the average score on that item divided by the highest number of points for any one alternative. Item difficulty is relevant for determining whether students have learned the concept being tested. It also plays an important role in the ability of an item to discriminate between students who know the tested material and those who

¹ Raw scores are those scores which are computed by scoring answer sheets against a ScorePak Key Sheet. Raw score names are EXAM1 through EXAM9, QUIZ1 through QUIZ9, MIDTRM1 through MIDTRM3, and FINAL. ScorePak cannot analyze scores taken from the bonus section of student answer sheets or computed from other scores, because such scores are not derived from individual items. Furthermore, separate analyses must be requested for different versions of the same exam.

do not. The item will have low discrimination if it is so difficult that almost everyone gets it wrong or guesses, or so easy that almost everyone gets it right.

To maximize item discrimination, desirable difficulty levels are slightly higher than midway between chance and perfect scores for the item. (The chance score for five-option questions, for example, is .20 because one-fifth of the students responding to the question could be expected to choose the correct option by guessing.) Ideal difficulty levels for multiple-choice items in terms of discrimination potential are:

Format	Ideal Difficulty
Five-response multiple-choice	70
Four-response multiple-choice	74
Three-response multiple-choice	77
True-false (two-response multiple-choice)	85

(from Lord, F.M. "The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties," *Psychometrika*, 1952, 18, 181-194.)

ScorePak® arbitrarily classifies item difficulty as "easy" if the index is 85% or above; "moderate" if it is between 51 and 84%; and "hard" if it is 50% or below.

Item Discrimination. Item discrimination refers to the ability of an item to differentiate among students on the basis of how well they know the material being tested. Various hand calculation procedures have traditionally been used to compare item responses to total test scores using high and low scoring groups of students. Computerized analyses provide more accurate assessment of the discrimination power of items because they take into account responses of all students rather than just high and low scoring groups.

The item discrimination index provided by ScorePak® is a Pearson Product Moment correlation² between student responses to a particular item and total scores on all other items on the test. This index is the equivalent of a point-biserial coefficient in this application. It provides an estimate of the degree to which an individual item is measuring the same thing as the rest of the items.

Because the discrimination index reflects the degree to which an item and the test as a whole are measuring a unitary ability or attribute, values of the coefficient will tend to be lower for tests measuring a wide range of content areas than for more homogeneous tests. Item discrimination indices must always be interpreted in the context of the type of test which is being analyzed. Items with low discrimination indices are often ambiguously worded and should be examined. Items with negative indices should be examined to determine why a negative value was obtained. For example, a negative value may indicate that the item was miskeyed, so that students who knew the material tended to choose an unkeyed, but correct, response option.

Tests with high internal consistency consist of items with mostly positive relationships with total test score. In practice, values of the discrimination index will seldom exceed .50 because of the differing shapes of item and total score distributions. ScorePak® classifies item discrimination as "good" if the index is above .30; "fair" if it is between .10 and .30; and "poor" if it is below .10.

Alternate Weight. This column shows the number of points given for each response alternative. For most tests, there will be one correct answer which will be given one point, but ScorePak® allows multiple correct alternatives, each of which may be assigned a different weight.

² A correlation is a statistic which indexes the degree of linear relationship between two variables. If the value of one variable is related to the value of another, they are said to be "correlated." In positive relationships, the value of one variable tends to be high when the value of the other is high, and low when the other is low. In negative relationships, the value of one variable tends to be high when the other is low, and vice versa. The possible values of correlation coefficients range from -1.00 to 1.00. The strength of the relationship is shown by the absolute value of the coefficient (that is, how large the number is, whether it is positive or negative). The sign indicates the direction of the relationship (whether positive or negative).

Means. The mean total test score (minus that item) is shown for students who selected each of the possible response alternatives. This information should be looked at in conjunction with the discrimination index; higher total test scores should be obtained by students choosing the correct, or most highly weighted alternative. Incorrect alternatives with relatively high means should be examined to determine why "better" students chose that particular alternative.

Frequencies and Distribution. The number and percentage of students who choose each alternative are reported. The bar graph on the right shows the percentage choosing each response. Frequently chosen wrong alternatives may indicate common misconceptions among the students.

Difficulty and Discrimination Distributions

At the end of the Item Analysis report, test items are listed according their degrees of difficulty (easy, medium, hard) and discrimination (good, fair, poor). These distributions provide a quick overview of the test, and can be used to identify items which are not performing well and which can perhaps be improved or discarded.

Test Statistics

Two statistics are provided to evaluate the performance of the test as a whole.

Reliability Coefficient. The reliability of a test refers to the extent to which the test is likely to produce consistent scores. The particular reliability coefficient computed by ScorePak® reflects three characteristics of the test:

1. The intercorrelations among the items -- the greater the relative number of positive relationships, and the stronger those relationships are, the greater the reliability. Item discrimination indices and the test's reliability coefficient are related in this regard.
2. The length of the test -- a test with more items will have a higher reliability, all other things being equal.
3. The content of the test -- generally, the more diverse the subject matter tested and the testing techniques used, the lower the reliability.

Reliability coefficients theoretically range in value from zero (no reliability) to 1.00 (perfect reliability). In practice, their approximate range is from .50 to .90 for about 95% of the classroom tests scored by ScorePak®.

High reliability means that the questions of a test tended to "pull together." Students who answered a given question correctly were more likely to answer other questions correctly. If a parallel test were developed by using similar items, the relative scores of students would show little change.

Low reliability means that the questions tended to be unrelated to each other in terms of who answered them correctly. The resulting test scores reflect peculiarities of the items or the testing situation more than students' knowledge of the subject matter.

As with many statistics, it is dangerous to interpret the magnitude of a reliability coefficient out of context. High reliability should be demanded in situations in which a single test score is used to make major decisions, such as professional licensure examinations. Because classroom examinations are typically combined with other scores to determine grades, the standards for a single test need not be as stringent. The following general guidelines can be used to interpret reliability coefficients for classroom exams:

Reliability	Interpretation
.90 and above	Excellent reliability; at the level of the best standardized tests
.80 - .90	Very good for a classroom test
.70 - .80	Good for a classroom test; in the range of most. There are probably a few items which could be improved.
.60 - .70	Somewhat low. This test needs to be supplemented by other measures (e.g., more tests) to determine grades. There are probably some items which could be improved.
.50 - .60	Suggests need for revision of test, unless it is quite short (ten or fewer items). The test definitely needs to be supplemented by other measures (e.g., more tests) for grading.
.50 or below	Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.

The measure of reliability used by ScorePak® is Cronbach's Alpha. This is the general form of the more commonly reported KR-20 and can be applied to tests composed of items with different numbers of points given for different response alternatives. When coefficient alpha is applied to tests in which each item has only one correct answer and all correct answers are worth the same number of points, the resulting coefficient is identical to KR-20.

(Further discussion of test reliability can be found in J. C. Nunnally, *Psychometric Theory*. New York: McGraw-Hill, 1967, pp. 172-235, see especially formulas 6-26, p. 196.)

Standard Error of Measurement. The standard error of measurement is directly related to the reliability of the test. It is an index of the amount of variability in an individual student's performance due to random measurement error. If it were possible to administer an infinite number of parallel tests, a student's score would be expected to change from one administration to the next due to a number of factors. For each student, the scores would form a "normal" (bell-shaped) distribution. The mean of the distribution is assumed to be the student's "true score," and reflects what he or she "really" knows about the subject. The standard deviation of the distribution is called the standard error of measurement and reflects the amount of change in the student's score which could be expected from one test administration to another.

Whereas the reliability of a test always varies between 0.00 and 1.00, the standard error of measurement is expressed in the same scale as the test scores. For example, multiplying all test scores by a constant will multiply the standard error of measurement by that same constant, but will leave the reliability coefficient unchanged.

A general rule of thumb to predict the amount of change which can be expected in individual test scores is to multiply the standard error of measurement by 1.5. Only rarely would one expect a student's score to increase or decrease by more than that amount between two such similar tests. The smaller the standard error of measurement, the more accurate the measurement provided by the test.

(Further discussion of the standard error of measurement can be found in J. C. Nunnally, *Psychometric Theory*. New York: McGraw-Hill, 1967, pp.172-235, see especially formulas 6-34, p. 201.)

A CAUTION in Interpreting Item Analysis Results

Each of the various item statistics provided by ScorePak® provides information which can be used to improve individual test items and to increase the quality of the test as a whole. Such statistics must always be interpreted in the context of the type of test given and the individuals being tested. W. A. Mehrens and I. J. Lehmann provide the following set of cautions in using item analysis results (Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, 1973, 333-334):

1. Item analysis data are not synonymous with item validity. An external criterion is required to accurately judge the validity of test items. By using the internal criterion of total test score, item analyses reflect internal consistency of items rather than validity.
2. The discrimination index is not always a measure of item quality. There is a variety of reasons an item may have low discriminating power: (a) extremely difficult or easy items will have low ability to discriminate but such items are often needed to adequately sample course content and objectives; (b) an item may show low discrimination if the test measures many different content areas and cognitive skills. For example, if the majority of the test measures "knowledge of facts," then an item assessing "ability to apply principles" may have a low correlation with total test score, yet both types of items are needed to measure attainment of course objectives.
3. Item analysis data are tentative. Such data are influenced by the type and number of students being tested, instructional procedures employed, and chance errors. If repeated use of items is possible, statistics should be recorded for each administration of each item.