ANALYSIS REPORT BRIEF GUIDE

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.

Item analysis report provides the following item information:

Item Number

This is the question number taken from the student answer sheet.

Mean Score

The mean is the "average" student response to an item. It is computed by adding up the number of points earned by all students on 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 higher the value of the standard deviation, the better the test is discriminating among student performance levels.

Median score

This is the raw score point that divides the raw score distribution in half; 50% of the scores fall above the median and 50% fall below.

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 zero to 100; the higher the value, the easier the question.

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 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	0.70
Four-response multiple-choice	0.74
Three-response multiple-choice	0.77
True-false (two-response multiple-choice)	0.85

An arbitrary classification for item difficulty:

"very easy"	> 0.95
"easy"	≥ 0.85
"moderate"	0.51-0.84
"hard"	≤ 0.50
"very hard"	< 0.30

Item Discrimination

Item discrimination refers to the ability of an item to differentiate among students based on 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 is a Pearson Product Moment correlation between student responses to a particular item and total scores on all other items on the test. This index is the equivalent of point-biserial coefficient. It provides an estimate of the degree to which an individual item is measuring the same thing as the rest of the items.

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 mis-keyed, 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 0.5 because of the differing shapes of item and total score distributions.

An arbitrary classification for item discrimination index:

"good"	> 0.3
"fair"	0.1-0.3
"poor"	**< 0.1**

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. 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

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

Number of Items

Number of items on the test.

Mean Score

Arithmetic average; the sum of all scores divided by the number of scores.

Median Score

The raw score point that divides the raw score distribution in half; 50% of the scores fall above the median and 50% fall below.

Standard Deviation

Measure of the spread or variability of the score distribution. The higher the value of the standard deviation, the better the test is discriminating among student performance levels.

Reliability (KR-20)

Is an estimate of test reliability indicating the internal consistency of the test. The range of the reliability is from 0.00 to 1.00. A reliability of .70 or better is desirable for classroom tests.

The reliability of a test refers to the extent to which the test is likely to produce consistent scores. The measure of reliability used 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.

Reliability (KR-21)

When item difficulties are approximately equal, is an estimate of test reliability indicating the internal consistency of the test. The range of the reliability is from 0.00 to 1.00. A reliability of .70 or better is desirable for classroom test.

≥ .90	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	Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.

S.E. of Measurement

The accuracy of measurement expressed in the test score scale. The larger the standard error, the less precise the measure of student achievement. 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 zero and 1.00, the standard error of measurement is expressed in the same scale as the test scores.

A Caution in Interpreting Item Analysis Results

Each of the various item statistics 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):

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.

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.

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.

References

Center for Innovation in Teaching & Learning, University of Illinois.

Office of Educational Assessment, University of Washington.

Written by

Dr. Mohamed Bahr

Deputy Head of Exam Center

Head of Assessment Committee