Interpreting the Item Analysis Report

This document is prepared to help instructors interpret the statistics reported on the Item Analysis Report and improve the effectiveness of test items and the validity of test scores.

Correct Responses as a Percentage of the Total Group: The proportion of students answering an item correctly indicates the difficulty level of the item. The more students got the item right, the less difficult the item was. Optimally, an item will encourage a widespread distribution of scores if its difficulty index is approximately 0.5 (i.e. 50% of the students got it right).

Percentage Range	Difficulty Index	Interpretation
75%-100%	0.75 – 1.0	Easy
26%-74%	0.25 – 0.75	Average
0-25%	0.25 or below	Hard

Correct Responses as a Percentage of the Upper/Lower 27% of Group: The upper and lower 27% rule is commonly used in item analysis based on Kelley's (1939) derivation. The difference between the correct responses as a percentage of the upper 27% and lower 27% of the total group can tell us whether an item has discriminated the high scorers and low scorers on the test. Discrimination Index (Point Biserial) is a much more robust statistic to examine an item's discrimination power.

Discrimination Index (Point Biserial): The discrimination index is a statistic which indicates the extent to which an item has discriminated between the high scorers and low scorers on the test. The index is represented as a fraction and varies between -1 to 1. Optimally an item should have a positive discrimination index of at least 0.2, which indicates that high scorers have a high probability of answering correctly and low scorers have a low probability of answering correctly. Items with negative indices should be examined to determine whether the item was flawed or miskeyed.

Discrimination Index	Interpretation
0.30 and above	Good
0.10 – 0.30	Fair
Equal to 0	No discrimination. All students got the item right.
Negative	Poor. The item was flawed or miskeyed.

Mean score: The mean score is the average of the test scores for the class. It gives a rough idea of how students performed as a whole. When we compare a student's score to the mean, we can say that that student did less well or did better than the class.

Median score: The median score is the middle test score when all the test scores are arranged in numerical order. Unlike the mean score, the median score is not strongly affected by extreme test scores (outliers). Therefore, it is a better estimate of students' average performance when some extreme test scores exist.

Non-Distractors: Non-distractors are item options that are not chosen by any student. Therefore, they do not provide any information to distinguish different levels of student performance. When an item has too many non-distractors, it needs to be revisited and possibly revised.

Percentile: The percentile score is a transformation of the raw score. It is used to demonstrate how much percentage of students in the class scored below a particular corresponding raw score. For example, Jane's test raw score is 86, and her percentile score is 94%. We can say that Jane's test score 86 on this test was higher than the scores of 94% of the students in the class.

Reliability Coefficient (KR20): Test reliability is an indication of the consistency of the test. It is not an index of quality ("Is this test a good measure of …?"), but of relative reproducibility ("How repeatable is this test?"). Test reliability tells us how likely it is that a student would obtain the same score when he/she takes the test again. This index can hold a value between 0 and 1. Higher test reliability indicates that the test measures whatever it measures in a consistent manner.

Reliability Coefficient	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. 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.

Note: The guidelines used to interpret reliability coefficients for classroom exams are adapted from http://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html.

Standard deviation: The standard deviation is a statistic which tells us how widely the scores are spread from the mean. A large standard deviation means that there is much variability in the test scores of the group (i.e. students performed quite differently on the test). A small standard deviation means that there is little variability amongst the scores (i.e. students performed quite similarly on the test).

Reference:

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17-24.

Need assistance in interpreting your Test Item Analysis Report, or suggestions on how to improve validity and reliability of your test? Please feel free to contact The Faculty Center by phone at 2-2783 or through email at facultycenter@stonybrook.edu.