Item Difficulty

Perhaps "item difficulty" should have been named "item easiness;" it expresses the proportion or percentage of students who answered the item correctly. Item difficulty can range from 0.0 (none of the students answered the item correctly) to 1.0 (all of the students answered the item correctly). Experts recommend that the average level of difficulty for a four-option multiple choice test should be between 60% and 80%; an average level of difficulty within this range can be obtained, of course, when the difficulty of individual items falls outside of this range. If an item has a low difficulty value, say, less than .25, there are several possible causes: the item may have been miskeyed; the item may be too challenging relative to the overall level of ability of the class; the item may be ambiguous or not written clearly; there may be more than one correct answer. Further insight into the cause of a low difficulty value can often be gained by examining the percentage of students who chose each response option. For example, when a high percentage of students chose a single option other than the one that is keyed as correct, it is advisable to check whether a mistake was made on the answer key.

Item Discrimination

The point-biserial correlation is an index of item discrimination, i.e., how well the item serves to discriminate between students with higher and lower levels of knowledge. The point-biserial correlation reflects the degree of relationship between scores on the item -- 0=incorrect, 1=correct -- and total test scores. Thus the point-biserial will be positive if better students answered the item correctly more frequently than poorer students did, and negative if the opposite occurred. A negative point-biserial is denoted by a minus sign in front of the value.

The value of a positive point-biserial discrimination index can range between 0 and 1; the closer the value is to 1, the better the discrimination. (The value of a negative point-biserial discrimination index can range between -1 and 0, but positive values are desirable). Item discrimination is greatly influenced by item difficulty. Items with a difficulty of either 0 or 1 will always have a discrimination index of 0, and item discrimination is maximized when item difficulty is close to .5. As a general rule, point-biserial values of .20 and above are considered to be desirable.

Items with negative discrimination values should be reviewed. A negative discrimination value, like a low difficulty value, may occur as a result of several possible causes: a miskeyed item, an item that is ambiguous, or an item that is misleading. As was previously said with respect to low difficulty values, further insight into the cause of negative discrimination can often be gained by examining the distribution of student responses. In small classes negative values close to zero are not necessarily reason for concern; they may be caused by one good student answering the item incorrectly or one poor student answering the item correctly.

KR 20

KR 20 is an index of the internal consistency of the test. "Internal consistency" refers to consistency of students' responses across the items on the test. KR 20 can be thought of as a measure of the extent to which the items on a test provide consistent information about a students' level of knowledge of the content assessed by the test. Assuming that all the items on a test relate to a single content domain, we would expect students with a very high level of knowledge of the domain to answer most items correctly and students with a very low level of knowledge of the domain to answer most items incorrectly.

The value of KR 20 can range from 0 to 1, with numbers closer to 1 reflecting greater internal consistency. (The value of KR 20 may be negative under certain circumstances, but this is a rare occurrence). What are acceptable values of KR 20? There is no single answer to this question. As a general rule, values of KR 20 for professionally developed and widely administered tests such as the SAT or GRE are expected to be greater than or equal to .80. Values of KR 20 for tests developed by instructors are not held to the same standard; one rule of thumb states that values greater than or equal to .70 are acceptable. Values of KR 20 for tests that assess several content areas or topics are expected to be lower than values of KR 20 than tests that assess a single content area.

When interpreting the value of KR 20 two additional factors should be considered: the size of the class and the extent of variability in students' knowledge. Values of KR 20 for small classes (say, less than 15 students) should be interpreted with caution since the observed value may be considerably different than the value that would be obtained if the test were administered to a larger sample. When a class is made up of students who are similar to each other in their level of ability, i.e., the range of test scores is small, the observed value of KR 20 will generally be lower than the value that would be obtained if the test were administered to a more diverse sample of students.