Source: UNIVERSITY OF WISCONSIN OSHKOSH

http://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/item-difficulty

OPTIMAL ITEM DIFFICULTY (PERCENT CORRECT)

"Item difficulty" is the percentage of the total group that got the item correct. Item difficulty is important because it reveals whether an item is too easy or too hard. In either case, the item may add to the unreliability of the test because it does not aid in differentiating between those students who know the material and those who do not. For example, an item answered correctly by everyone does nothing to aid in the assignment of grades. The same is true for items that no one answers correctly.

The **optimal item difficulty depends on the question-type and on the number of possible distractors.** Many test experts believe that for a maximum discrimination between high and low achievers, the optimal levels (adjusting for guessing) are:

Question type/number of distractors	Recommended item Difficulty
True and false (2)	.75
Multiple choice, three distractors	.67
Multiple choice, four distractors	.63
Multiple choice, five distractors	.60

Items with difficulties less than 30 percent or more than 90 percent definitely need attention. Such items should either be revised or replaced. An exception might be at the beginning of a test where easier items (90 percent or higher) may be desirable.

Source: UNIVERSITY OF WISCONSIN OSHKOSH

http://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/item-i

ITEM DISCRIMINATION I: DISCRIMINATION INDEX

The single best measure of the effectiveness of an item is its ability to separate students who vary in their degree of knowledge of the material tested and their ability to use it. If one group of students has mastered the material and the other group had not, a larger portion of the former group should be expected to correctly answer a test item. Item discrimination is the difference between the "percentage correct" for these two groups.

Item discrimination can be calculated by ranking the students according to total score and then selecting the top 27 percent and the lowest 27 percent in terms of total score. For each item, the percentage of students in the upper and lower groups answering correctly is calculated. The difference is one measure of item discrimination (ITEM DISCRIMINATION INDEX). The formula is:

IDI = (Upper Group Correct Answers – Lower Group Correct Answers)/27% examinees

The maximum item discrimination difference is 100 percent. This would occur if all those in the upper group answered correctly and all those in the lower group answered incorrectly.

Zero discrimination occurs when equal numbers in both groups answer correctly.

Negative discrimination, a highly undesirable condition, occurs when more students in the lower group than the upper group answer correctly.

The following levels may be used as a guideline for acceptable items.

IDIs	Interpretation
<0	**Unacceptable – check item for error**
0.0 – 0.24	**Usually unacceptable – might be approved**
0.25 – 0.39	**Good item**
0.40 – 1.0	**Excellent item**

Source: UNIVERSITY OF WISCONSIN OSHKOSH

http://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/item-ii

ITEM DISCRIMINATION II POINT BISERIAL CORRELATION (PBC)

The point biserial correlation (PBC) measures the correlation between the correct answer (viewed as 1 = right and 0 = wrong) on an item and the total test score of **all students**. The PBC is sometimes preferred because it identifies items that correctly discriminate between high and low groups, as **defined by the test as a whole instead of the upper and lower 27 percent of a group**.

PBC can generate a substantially different measurement of item discrimination than the simple item discrimination difference (ITEM DISCRIMINATION INDEX) described above. **Often, however, the measures are in close agreement.**

Generally, the higher the PBC the better the item discrimination, and thus, the effectiveness of the item. The following criteria may be used to evaluate test items.

PBC	Interpretation
.30 and above	**Very good items**
.20 to .29	**Reasonably good items, but subject to improvement**
.10 to .19	**Marginal items, usually needing improvement**
.00 to .09	**Poor items, to be rejected or revised**

Source: UNIVERSITY OF WISCONSIN OSHKOSH

http://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/distractors-and-effectiveness

DISTRACTORS & EFFECTIVENESS

Although Item Discrimination statistics measure important characteristics about test item effectiveness, they don't reveal much about the appropriateness of item distractors. By looking at the pattern of responses to distractors, teachers can often determine how to improve the test.

The effectiveness of a multiple-choice question is heavily dependent on its distractors. **If two distractors in a four-choice item are implausible, the question becomes, in effect, a true false item.** It is, therefore, important for teachers to observe how many students select each distractor and to revise those that draw little or no attention. Use of "all of the above" and "none of the above" is generally discouraged.

Source: UNIVERSITY OF WISCONSIN OSHKOSH

http://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/reliability-validity

RELIABILITY & VALIDITY

The importance of a test achieving a reasonable level of reliability and validity cannot be overemphasized. To the extent a test lacks reliability, the meaning of individual scores is ambiguous. A score of 80, say, may be no different than a score of 70 or 90 in terms of what a student knows, as measured by the test. If a test is not reliable, it is not valid.

Reliability of a Test

Despite differences between the format and construction of various tests, there are two standards by which tests (as compared to items) are assessed. These two standards are reliability and validity.

Reliability refers to the consistency of test scores; how consistent a particular student's test scores are from one testing to another. In theory, if test A is administered to Class X, and one week later is administered again to the same class, individual scores should be about the same both times (assuming unchanging conditions for both sessions, including familiarity with the test). If the students received radically different scores the second time, the test would have low reliability. Seldom, however, does a teacher administer a test to the same students more than once, so the reliability coefficient must be calculated a different way. Conceptually, this is done by dividing a homogeneous test into two parts (usually even and odd items) and treating them as two tests administered at one sitting. The calculation of the reliability coefficient, in effect, compares all possible halves of the test to all other possible halves.

One of the best estimates of reliability of test scores from a single administration of a test is provided by the **Kuder-Richardson Formula 20 (KR20)**. For good classroom tests, the reliability coefficients should be .70 or higher.

To increase the likelihood of obtaining higher reliability, a teacher can:

(a) Increase the length of the test.

(b) Include questions that measure higher, more complex levels of learning, and include questions with a range of difficulty with most questions in the middle range.

(c) If one or more essay questions are included on the test, grade them as objectively as possible.

Validity of a Test

Content or curricular validity is generally used to assess whether a classroom test is measuring what it is supposed to measure. For example, a test is said to have content validity if it closely parallels the material which has been taught and the thinking skills that have been important in the course. Whereas reliability is expressed as a quantitative measure (e.g., .87 reliability), content validity is obtained through a rational or logical analysis of the test. That is, one logically compares the test content with the course content and determines how well the former represents the latter.

A quantitative method of assessing test validity is to examine each test item. This is accomplished by reviewing the discrimination (IDI) of each item. If an item has a discrimination measure of 25 percent or higher, it is said to have validity, that is, it is doing what it is suppose to be doing – discriminating between those that are knowledgeable and those that are not knowledgeable.