

Suggestions for using Digitek statistics to improve multiple-choice test items

Statistic	What it means	What it may suggest
Percent correct: Top/bottom third	For each test item, reports the percent of students ranking in the top third of the class on total exam score who got this item right; and the percent of students in the bottom third on total score who got this item right.	This is a quick indication of how each item “performed” on the exam. Generally, more students from the top third should get an item right than students from the bottom third. Easy items will have high percents for both groups; very difficult, confusing, or mis-keyed items may have low percents for both.
“P” values for each response option	For each response option for each test item, P indicates the percent of students who chose that response option.	P should generally be highest for the correct response option. Good distracters (incorrect answer options) should be selected by some students. Distracters with very few or no “takers” might be considered for replacement since they are not helping the test discriminate between students who have and have not mastered the information.
“R” values for each response option	For each response option for each test item, R indicates the correlation between the selection of that response option and total score on the test. (The value of R can vary from –1.0 to 1.0). Technically, R is called the <i>discrimination index</i> for the item.	For the correct answer option (indicated by *) the correlation should usually be a positive number from about .30 to .90. This shows that students who got the item right tended to score well on the overall test. The distracters should have negative or near-zero Rs, since choosing them does not contribute to overall score. Correct answers with negative or near-zero Rs may indicate confusing, ambiguous or mis-keyed items.
Overall indices of test reliability (Kuder-Richardson and Spearman-Brown)	These statistics, which generally range from 0 to 1.0, indicate the internal consistency of the test or the extent to which the items work together to measure the same underlying construct.	These statistics provide, in a single number, similar information as the individual item discrimination indices (Rs). If the R values for most items are high positive numbers, the overall test reliability indices should also be high. If many R values are low, the reliability indices will also probably be low. Generally, the more items on a test the higher the reliability index is likely to be. Also, note that exams with mostly very high scores or mostly very low scores tend to have lower reliability indices.