

Test and Item Analysis

This Helpful Teaching Hints explains two measurement concepts that will help you analyze your medical student examinations: (1) test reliability and (2) test item discrimination.

Test Reliability

When you scan and score an examination in the Office of Medical Education, the Remark software reports two measures of test reliability: **Kuder-Richardson 20** (KR-20) and Coefficient Alpha. The latter is best used with surveys or attitude measures. The KR-20 index is the appropriate index of test reliability for multiple-choice examinations.

What does the KR-20 measure?

The KR-20 is a measure of internal consistency reliability or how well your exam measures a single cognitive factor. If you administer an Embryology exam, you hope all test items relate to this broad construct. Similarly, an Obstetrics/Gynecology test is designed to measure this medical specialty.

How do you interpret the KR-20 value?

The KR-20 formula includes (1) the number of test items on the exam, (2) student performance on every test item, and (3) the variance (standard deviation squared) for the set of student test scores. The index ranges from 0.00 to 1.00. A value close to 0.00 means you are measuring many unknown factors but not what you intended to measure. You are close to measuring a single factor when your KR-20 is near 1.00. Most importantly, we can be confident that an exam with a high KR-20 has yielded student scores that are **reliable** (i.e., reproducible or consistent; or as psychometricians say, the true score). A medical school test should have a KR-20 of 0.60 or better to be acceptable.

How can I improve my test reliability?

Measurement experts have long known the following:

1. In general, **longer tests** (more test items) give more reliable scores. This is intuitive. The instructor is more likely to measure your true score if a test has 50 items rather than, say, 10 items.
2. In general, the more **heterogeneous** the group of students, the higher the reliability. Some think that medical students are a homogeneous group of high achievers, but the experienced teacher knows differently. A class of medical students is sufficiently variable in knowledge, and their test scores, to support reasonable test reliability.

3. Related to group heterogeneity in test scores is **item difficulty**. Reliability is maximized when items are answered correctly by half the students. However, this is undesirable; we do not want a mean score of 50%. In practical terms, an exam where about 70-80 percent of students answer test items correctly will yield reliable scores.

Test Item Discrimination

When you scan and score an examination in the Office of Medical Education, the Remark software reports an index of test item discrimination, the **point biserial correlation**.

What does the point biserial measure?

A correlation is a statistic that quantifies the relationship between two variables. The scale of measurement for the two variables determines which specific correlation is appropriate. Since we want to correlate a test item, a **categorical** variable (i.e., the student either answered the test item correctly or incorrectly), with a **continuous** variable (i.e., percent score on the examination), we need the correlation index for a categorical variable and a continuous variable. This is the point biserial correlation.

How do you interpret a point biserial correlation?

As with all correlation indices, the point biserial ranges from -1.00 to $+1.00$. A positive point biserial tells us that those scoring higher on the exam were more likely to answer the test item correctly (i.e., the item “discriminates” between high-scoring and low-scoring students). Conversely, a negative point biserial says that high scorers on the exam answered the test item incorrectly more frequently than low scorers. We prefer the former rather than the latter. A negative point biserial suggests an unpleasant explanation – e.g., the item was keyed incorrectly, the item was poorly constructed or misleading, the content of the item was inadequately taught.

What is a desirable point biserial correlation for a test item?

The higher, the better. As a general rule, $+0.20$ is desirable. However, there is an interaction between item discrimination and item difficulty, and you should be aware of two principles:

1. very easy or very difficult test items have little discrimination
2. items of moderate difficulty (60% to 80% answering correctly) generally are more discriminating.

Examine the test item results below.

Response ITEM #1	Correct	Percent Choosing	Difficulty	Point Biserial Correlation
A	Correct	.72	.28	.40
B		.03		
C		.04		
D		.09		
E		.12		
Response ITEM #2	Correct	Percent Choosing	Difficulty	Point Biserial Correlation
A	Correct	.72	.28	.04
B		.03		
C		.04		
D		.09		
E		.12		
Response ITEM #3	Correct	Percent Choosing	Difficulty	Point Biserial Correlation
A		.01		
B		.00		
C		.01		
D	Correct	.98	.02	.00
E		.00		
Response ITEM #4	Correct	Percent Choosing	Difficulty	Point Biserial Correlation
A		.02		
B		.15		
C		.05		
D	Correct	.70	.30	- .19
E		.08		

Item #1 and Item #2 both were correctly answered by 72% of the class. We are pleased with Item #1 because of its hefty point biserial of .40 (i.e., high-scoring students much more likely to choose the correct answer). We are suspect of Item #2. A point biserial of .04 on a test item answered incorrectly by 28% of the students suggests inadequate discrimination and a flawed item.

A point biserial of .00 for Item #3 is not surprising given that nearly every student (98%) answered the item correctly. Incidentally, it is satisfactory to have some test items that all or nearly answer correctly. Some concepts are so important that we expect few, if anyone, to not have this knowledge.

Finally, Item #4 is likely flawed. Difficulty was in the sweet spot – 70% answering correctly – but for some reason the high-scoring students were notably more likely to answer incorrectly. You need to investigate Item #4 carefully.