Using Remark Statistics for Test Reliability and Item Analysis

- Shannon Tucker, Director of Instructional Technology, University of Maryland School of Pharmacy

Test Reliability

There are numerous indexes that may be used to assess the internal consistency of an assessment. Currently, the most widely used measure of reliability is Cronbach's Alpha (also known as the Coefficient Alpha)[1]. However, you will notice that test statistics included with every scored assessment will include both the Kuder-Richardson Formula (KR-20) and the Coefficient Alpha (Cronbach's Alpha). The Coefficient Alpha is most often used on instruments where items are not scored as right or wrong[2]. KR-20 is a special case of Cronbach's alpha specifically for ordinal dichotomies to evaluate how consistent student responses are among questions on an assessment[3]. In laymen's terms KR-20 best measures how well your exam measures a subject (a single cognitive factor) and the Coefficient Alpha best measures surveys or attitude data.

Interpreting KR-20[4]

KR-20 formula includes:

1. Number of test items on the exam
2. Student performance on every test item
3. Variance

Index Range: 0.00-1.00

Values near 0.00: Measuring many unknown factors, but not what you intended to measure

Values near 1.00: Close to measuring a single factor

Summary: An exam with a high KR-20 yields reliable student scores (consistent/true score)

---

[1] Streiner, David L. "Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency." Journal of Personality Assessment 80(1) (2003): 99-103.

[2] Reliability. Del Siegle Faculty Web Site University of Connecticut. < http://www.gifted.uconn.edu/siegle/research/Instrument%20Reliability%20and%20Validity/Reliability.htm> 19 September 2007.

[3] Scales and Standard Measures. North Carolina State University. 19 September 2007.

[4] Test and Item Analysis. Tulane University Office of Medical Education. < http://www.som.tulane.edu/ome/helpful_hints/test_analysis.pdf> 19 September 2007.

How others use KR-20: Tulane University Office of Medical Education recommends a KR-20 score of 0.60 or larger to be acceptable.

Item Analysis

Conducting an item analysis following an administration of your assessment is important to identify any questions that are not performing well due to inappropriate difficulty, scoring error, or other factors. When conducting an item analysis the item difficulty, item discrimination, and distractor quality should all be considered.

Item Difficulty (p-value)

Item Difficulty is a measure of the proportion of students/subjects who have answered an item correctly and is most commonly referred to as the *p-value*.

Index Range: 0.00-1.00

Values near 0.00: A greater proportion of students/subjects responded to the item correctly (more difficult)

Values near 1.00: A greater proportion of students/subjects responded to the item correctly (easier)

Summary: The p-value will report item difficulty related to your assessed population.

How others use the p-value: Consulting company Professional Testing suggests item difficulty for criterion-referenced tests (CRTs), with their emphasis on mastery-testing, many items on an exam form will have p-values of .9 or above. Norm-referenced tests (NRTs), are designed to be harder overall and to spread out the examinees' scores. Thus, many of the items on an NRT will have difficulty indexes between 0.4 and 0.6.[5]

Item Discrimination

There are several indexes that successfully compute item discrimination. While the discrimination index is a popular and valid measure of item quality[6], this index is not included in as a part of the Remark reported item statistics. Instead Remark provides the *Point Biserial Correlation*.

Point Biserial Correlation

The Point Biserial Correlation quantifies the relationship between a student/subject's score (correct or incorrect) and the overall assessment score.

Index Range: -1.00 **- +**1.00

Values Near -1.00: High scorers answered the item incorrectly more frequently than low scorers.

Values Near +1.00: High scorers answered the item correctly more frequently than low scorers.

---

[5] Step 9. Conduct the Item Analysis. Building High Quality Examination Programs – Professional Testing. 2005.

[6] Pyrczak, Fred. "Validity of the Discrimination Index as a Measure of Item Quality." Journal of Educational Measurement. 10(3) (1973):227-231.

Summary: A negative value indicates an item may have been misleading, keyed incorrectly, or the content was inadequately covered.

How others use the point biserial correlation: Tulane University Office of Medical Education suggest to fauclty that a score of +0.20 is desirable[7].

They also suggest that there is an interaction between the item discrimination and item difficulty that should be considered by faculty:

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

Sample results from Tulane University Office of Medical Education using the p-value with the point biserial correlation can be found at:

http://www.som.tulane.edu/ome/helpful_hints/test_analysis.pdf

Distractor Analysis

Unfortunately, neither item difficulty or item discrimination account for incorrect response options (distracters). Distractor analysis will assist individuals with addressing performance issues associated with incorrect options. On a well-designed multiple choice item, high scoring students/subjects should select the correct option even from highly plausible distractors[8]. Those who are ill-prepared should select randomly from available distractors. In this scenario, the item would be a good discriminator of knowledge and should be considered for future assessments. In other scenarios, a distractor analysis may reveal an item that was mis-keyed, contained a proofreading error, or contains a distractor that appears plausible even by those that scored well on an assessment.

To be effective incorrect options should be plausible and incorrect without ambiguity. Therefore distractor analysis examines the proportion of students/subjects who selected each of the response options. For the correct response, this proportion is equivalent to the item p-value, or item difficulty[9]. If all response option proportions are summarized they will add up to 1.0 or 100% of student/subject selections. Reviewing the percentage of students/subjects who have responded to each response option will help you assess if there are issues present in an item's distractors.

Locating Distractor Statistics

To make distractor analysis easier, Remark returns a separate item analysis report specifically for distractor analysis (figure 2). Along with the label, value, weight, and frequency the item was selected, each question item analysis will also include the percent respondents selected an option and its corresponding point biserial correlation for all distractors in addition to the correct answer.

---

[7] Test and Item Analysis. Tulane University Office of Medical Education. < http://www.som.tulane.edu/ome/helpful_hints/test_analysis.pdf> 19 September 2007.

[8] Zurawski, Raymond M. Making the Most of Exams: Procedures for Item Analysis. National Teaching & Learning Forum. 7(6). 1998 . <http://www.ntlf.com/html/pi/9811/v7n6smpl.pdf>. 20 September 2007.

[9] Step 9. Conduct the Item Analysis. Building High Quality Examination Programs – Professional Testing. 2005.

Item Analysis: #1

Label	Value	Weight	Frequency	Percent	Point Biserial
A	1	1	19	95.00	0.57
B	2	0	0	0.00	-
C	3	0	1	5.00	-0.57
D	4	0	0	0.00	-
E	5	0	0	0.00	-
Total			20	100.00	

Figure 1: Distractor Analysis

Sample Item Analysis[10]

Good Item

P-Value: 0.72

Point Biserial: +0.22

Items	Frequency	Percent	Point Biserial
A (correct)	241	72.15	+0.22
B	9	2.70	-0.02
C	3	0.89	-0.10
D	11	3.30	-0.06
E	70	20.96	-0.19
Total	334	100	

This item is good because the point biserial correlation for the correct answer is above 0.2 and is higher than the same value for the other distractors.

Fair Item

P-Value: 0.39

Point Biserial: +0.12

Items	Frequency	Percent	Point Biserial
A	13	3.89	-0.18
B	87	26.05	-0.03
C	40	11.98	-0.10
D (correct)	130	38.92	+0.12
E	64	19.16	+0.05
Total	334	100	

[10] What Item Analysis Can Tell Us About Item Quality. Measurement Research. <http://measurementresearch.com/media/itemanalysis.pdf>. 20 September 2007.

While the point biserial correlation for this question is not above the desirable value 0.2, it is close to this value and is higher than the point biserial correlation for all distractors. Students/subjects answering incorrectly selected values from all distractors listed.

Poor Item

P-Value: 0.34

Point Biserial: -0.07

Items	Frequency	Percent	Point Biserial
A	2	0.60	+0.06
B	2	0.60	-0.16
C (correct)	113	33.83	-0.07
D	23	6.89	-0.02
E	194	58.08	+0.09
Total	334	100	

The point biserial correlation for this item is negative and lower than some of the other distractors identifying this as a poor question that should be evaluated for revision.