Center for Teaching Excellence
Hampton University
Teaching Matters

September/October 2016 Volume 11, Number 3

Featured Article:

An Introduction to Test Planning, Creating Test Items and Conducting Test Item Analysis, Part II

Dr. Simone Heyliger, Associate Professor of Pharmaceutical Sciences
Ms. Deborah Hudson, Assistant Professor of Pharmacy Practice

Developing quality test items is a necessary skill for college instructors. After the test items are developed and examinations are administered, it is important to appropriately analyze the items to identify inconsistencies and poor test questions. Part I of the article provided instructors with information on how to improve their writing of objectives and test items.

So now you've written the perfect objectives and the perfect test items.
What happens next?

ANSWER: Students "Bomb" Your Exam!

In most courses, tests are used as the main mechanism for measuring student performance and outcomes. Educators must be careful to *appropriately* measure what is being taught. Furthermore, there must be a balance between what is taught and what is assessed. Therefore, it is imperative to perform item analyses after administering an examination. It is important for identifying key concepts or subject matters that students may have had difficulty understanding. It is also important for identifying "keying" errors and/or poorly written examination questions. Item analysis is a method for reviewing items on a test, both qualitatively and statistically, to ensure that they all meet minimum quality-control criteria. The analysis is conducted after items have been administered and, when used appropriately, will identify poorly constructed test questions.

This article will review the following topics**:**

1. Simple descriptive statistics
    a. Measures of the center (mean/median/mode)
    b. Measures of dispersion (standard deviation/variance)
2. Item difficulty
3. Item discrimination
4. Validity and reliability
5. Point-biserial correlation coefficient
6. Kuder-Richardson 20 (KR-20)
7. Analysis of test items

Simple Descriptive Statistics

Measures of the center. After you have graded your examinations and disseminated the results, students desire to know how they compare with their peers. A *measure of the center* is a central or representative value computed from a data set. Three basic measures of the center are the *mean, median* and the *mode.* There is also the midrange, but we don't focus on it as much -- however, since we have mentioned it, let's define it, too.

- **Mean (sample)** – the sum of all the measurements divided by number of measurements;
- **Median** – The middle value if n (number of scores) is odd. The average of the 2 middle values, if n (number of scores) is even;
- **Mode** – The most frequent value if it exists;
- **Midrange** – The average of the smallest and largest values.

The "mean," that we report to our students, represents the arithmetic mean or average.

How do you determine which one to use? Consider the following:

- **Mean** is a more natural measure of location, but it is oversensitive to extreme values;
- **Median** is not affected by very large or very small values;
- **Mode** is not always unique; it is the only measure of central tendency that can be used with nominal data (categories with no order);
- **Midrange** is sensitive to extreme values.

If you have extreme values among your test scores, the *median* is a better measure of the center.

Measures of dispersion. In addition to measures of the center, we often review and report measures of dispersion or spread. The measures of dispersion most often used are the *standard deviation* and the *variance.*

Generally, a small standard deviation means that the data points (or test scores in our discussion) tend to be close to the mean; conversely, a large, high standard deviation indicates that the data points (or test scores) are more spread out. The standard deviation of a sample, population, data set, etc., is the square root of its variance. Unlike the variance, the standard deviation is easier to interpret since it is expressed in the same units as the data.

Item Difficulty

Item difficulty is expressed as the proportion (or percentage) of respondents who correctly answer a particular test item (out of those who responded). Item difficulty can range from 0 to +1, where "0" indicates that the item was correctly answered by no one, and "+1" indicates that the item was answered correctly by everyone. Item difficulty provides a guide of the "easiness" of a test item. Tests that are too easy may lead to exaggerated scores; conversely, tests that are too hard can lead to reduced scores – both extremes may lead to inaccurate and/or incomplete assessments.

Range of difficulty indices are necessary when preparing examinations. The average difficulty level for a multiple choice test with four (4) options ranges between 60% to 80%. Reasons why an item may have a low difficulty index (less than 21%) include: 1) there is more than one correct response for the item, 2) the item is too challenging for the class, 3) the item is not clearly written, and 4) the item is miss-keyed. The following table provides a rubric for interpreting the difficulty index (Figure 1).

Interpretation of the Difficulty Index

Range	Difficulty Level
20 & below	Very difficult
21-40	Difficult
41-60	Average
61-80	Easy
81 & above	Very easy

(adapted from https://ppukmdotorg.files.wordpress.com/2015/04/interpretdiff.png)

Using Figure 1, let's revisit the test planning matrix originally presented in Part 1 (Table 1). Based on the difficulty index (percentage of people who correctly answered), the questions that correspond to Bloom's Level 1 (*remember?)*, have the highest difficulty indices (70% to 95%), indicating easier questions, whereas SILO2, which measures Bloom's Level 4, has a lower difficulty index of 40%. (*Remember: high difficulty index = easy; low difficulty index = hard)*

SILO2 could be re-examined, since only 40% of the students who answered the item provided a correct response. However, this result does not indicate that the items in SILO2 are problematic; it simply provides a measurement of difficulty that one can use for re-examination of

the test, especially if the overall test scores are low or as an indication that students need further assessments in order to master questions at higher Bloom levels.

Table 1. Test Planning Matrix

SILOS	% of Test	Test Item Strategy	No. Test Q per Item	Bloom Level	Q No. assoc. w/ SILO	% students got all correct
SILO1	**50%**	**MCQ**	**15 questions**	**Level 3 (Apply)**	**1-5, 16-20, 26-30**	**60**
		TF	**5**	**Level 1 (Remember)**	**6-10**	**95**
		Matching	**5**	**Level 1 (Remember)**	**11-15**	**80**
SILO2	**20%**	**MCQ Math ques.**	**10 questions**	**Level 4 (Analyze)**	**21-25, 31-35**	**40**
SILO3	**10%**	**TF**	**5**	**Level 1 (Remember)**	**36-40**	**70**
SILO4	**10%**	**MCQ Fill in the blanks**	**5**	**Level 2 (Understand)**	**41-45**	**60**

(adapted from http://orgs.bloomu.edu/tale/documents/OAE4_TestBlueprinting.pdf)

Item Discrimination

When reviewing test items, it is important to note that discrimination extends beyond viewing those who correctly answer the question. Test items should also discriminate between groups. A basic consideration when evaluating a test item is the degree to which the item discriminates between high achieving students (high test scorers) and low achieving students (low test scorers). The *index of discrimination* is really a measure of item quality; furthermore, it is also a measure of whether the test item is too difficult or too hard.

The higher the index, the more likely the high test scorers will answer the question correctly. Item discriminations of 0.50 or more are interpreted as excellent. "No discrimination" is indicated by a score of 0 or near 0. An item discrimination that is equal to 1.00 is interpreted as having perfect discrimination.

When developing measurement tools, one must also consider whether it is valid and reliable. Simply stated, the *validity* of a measurement tool is the degree to which the tool measures what it claims to measure. *Reliability* is the overall consistency of a measure. We've often heard of reliable scales – this concept is important because we want to obtain consistent results when using selected measures.

Point-biserial correlation coefficient. Another index of item discrimination is the point-biserial correlation coefficient. It is a measure of item reliability (consistency). The point-biserial correlation indicates that test takers who performed well on the examination also selected the correct response, so this is also a good discriminator between high-scoring and low-scoring students. It compares how well students perform on

each test item relative to their overall performance on the test. Therefore, students who perform well on the test overall should also perform well on the test item being reviewed.

The point-biserial correlation ranges from -1 to +1. Items with incorrect keys will show a negative point biserial correlation.

General Interpretation

- Very Good Item: **.30 and above**
- Reasonably Good: **.20 - .29**
- Marginal Item: **.09 - .19**
- Poor Item: **below .09**

- Point Biserial =0 or near 0 means no discrimination. All students got the item right. Items with strong discrimination power will be close to 1.0.

But wait...there's more!

Kuder-Richardson 20 (KR-20). KR-20 is a measure of internal consistency (reliability) and measures how consistent students' responses are across test items.

- It is influenced by:
  - The number of items on the test
  - The difficulty of each item on the test
  - The variance of the students' test scores
- Ranges from 0 to 1
- The higher the value, the stronger the relationship among the items on the test
- Value of at least 0.70 is desirable

The following table contains suggested guidelines when interpreting the KR-20 (Table 2).

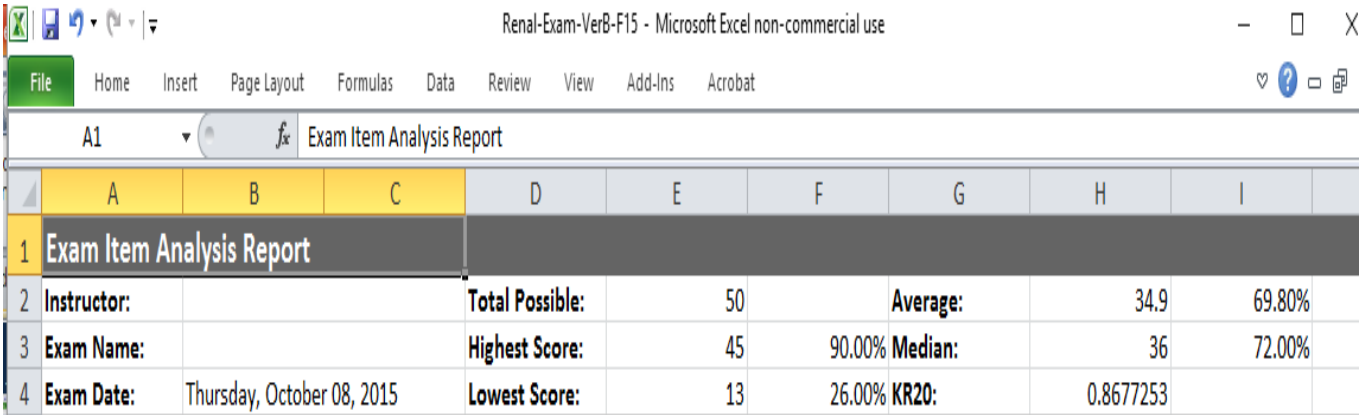
Table 2. Suggested KR-20 Guidelines

Reliability	Interpretation
.90 and above	**Excellent reliability; at the level of the best standardized tests**
.80 - .90	**Very good for a classroom test**
.70 - .80	**Good for a classroom test; in the range of most. There are probably a few items which could be improved.**
.60 - .70	**Somewhat low. This test needs to be supplemented by other measures (e.g., more tests) to determine grades. There are probably some items which could be improved.**
.50 - .60	**Suggests need for revision of test, unless it is quite short (ten or fewer items). The test definitely needs to be supplemented by other measures (e.g., more tests) for grading.**
.50 or below	**Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.**

(adapted from https://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html)

(The statistics discussed are generated by the Scantron System or Apperson DataLink System. Both software programs generate the point-biserial correlation, KR-20, mean, median, and standard deviation statistics. Scantron may provide additional statistics such as degree of kurtosis and measures of skewness. Apperson DataLink provides the discrimination index).

Figure 2. Exam Item Analysis Report



	A	B	C	D	E	F	G	H	I
1	**Exam Item Analysis Report**								
2	**Instructor:**			**Total Possible:**	50		**Average:**	34.9	69.80%
3	**Exam Name:**			**Highest Score:**	45	90.00%	**Median:**	36	72.00%
4	**Exam Date:**	Thursday, October 08, 2015		**Lowest Score:**	13	26.00%	**KR20:**	0.8677253	

As depicted in Figure 2, this examination generated a KR-20 = ***0.86. Based on the rubric, it is desirable, since it at least 0.70.***

Furthermore, we observe the mean (average) examination score is 69.80, compared to a median of 72. The mean and median are close but not equal. The median is a better measure of the center due to an extreme value (look at the lowest score).

Let us now apply the measures that we've discussed.

Analysis of Test Items

Figure 3 below lists the point-biserial correlation (pt. bis.) and discrimination (DISC) index statistics for several multiple choice questions. Using the information in the article, provide the best answers.

Figure 3. Multiple choice questions.

Q	A	B	C	D	E			Pt. bis.	DISC
Q23	*A (12, 41.38%)*	B (1, 3.45%)	C (13, 44.83%)	D (1, 3.45%)	E (2, 6.90%)			-0.07	0
Q24	A (6, 20.69%)	B (1, 3.45%)	*C (10, 34.48%)*	D (5, 17.24%)	E (7, 24.14%)			0.27	0.75
Q25	*A (24, 82.76%)*	B (0, 0.00%)	C (1, 3.45%)	D (2, 6.90%)	E (2, 6.90%)			0.33	0.29
Q26	A (6, 20.69%)	B (12, 41.38%)	C (1, 3.45%)	*D (9, 31.03%)*	E (1, 3.45%)			0.43	0.75
Q27	A (0, 0.00%)	*B (27, 93.10%)*	C (2, 6.90%)	D (0, 0.00%)	E (0, 0.00%)			0.13	0.14

1. Which Question most likely represents a keying error?

    A. Question 23

    B. Question 24

    C. Question 25

    D. Question 26

    E. Question 27

2. Which Question most likely represents an easy question?

    A. Question 23

    B. Question 24

    C. Question 25

    D. Question 26

    E. Question 27

ANSWERS:

1. Q23 – negative point-biserial, no discrimination; similar proportion of students chose A and C.

2. Q27 – overwhelming majority of students chose B; low point-biserial and low discrimination.

REFERENCES AND USEFUL LINKS

http://edassess.net/eacs/pointbiserial.aspx

http://edassess.net/eacs/pointbiserial.aspx

https://jcesom.marshall.edu/media/24104/Item-Stats-Point-Biserial.pdf

http://languagetesting.info/statistics/excel.html

http://orgs.bloomu.edu/tale/documents/OAE4_TestBlueprinting.pdf

http://www.omet.pitt.edu/

https://ppukm.org/2015/04/02/calculating-omr-indexes/

https://tulane.edu/som/ome/pd-exams-test-item-analysis-module.cfm

Daniels and Cross. Biostatistics: A Foundation for Analysis in the Health Sciences, 10th Edition. 2013