Handout for Webinar II: Rules of Thumb When Appraising Validity Data

Reliability (Correlation) Coefficients: How High Should They Be?

It depends on the purpose of measurement, but higher is always better. This is because the lower the reliability, the greater the measurement error in your scores.

- For high-stakes assessments (such as licensure, certification; results that impact resource utilization or lead to rewards or sanctions), the reliability should be **0.90** or above.
- For evaluation and research purposes, the reliability should be **0.80** or above.
- For moderate-stake assessments (such as end of course evaluations, year-end exams), the reliability should be in the **0.80** range and above.
- For low-stakes assessments (e.g., classroom tests, quizzes created by instructors) used mostly for teaching purposes, acceptable reliabilities could be in the **0.70** range and above.

Standard Error of Measurement: When to Worry?

Smaller is always better. The standard error of measurement (SEM) reflects the amount of error for a distribution of scores. You can think of the SEM as the amount that a person's obtained score from a given test would vary from his or her "true" score, if the test were given an infinite amount of times. The magnitude of the SEM can be best judged by looking at it in relationship to the scale range.

- A SEM of **1.0** on a scale of 1 to 5 is quite large (20% of the scale)
- A SEM of **1.0** on a scale of 1 to 50 is very acceptable (.02% of the scale).

Confidence Intervals: How Confident Do You Need to Be?

In laymen terms, confidence intervals represent the range of scores within which we can say the person's true score most probably lies, plus or minus some margin of error. The calculation of confidence intervals depends on how narrow or wide you set the score range, and the degree of confidence you aspire to (do you want to be 85% confident? 99% confident?). So the trick is to setting those parameters. It is pretty easy to be 95% confident that a mean sample score on a 10-point test would lie between 2.0 and 8.0 in the population of test takers; it is much more difficult to be 95% confident that it would lie between 7.0 and 8.0. The confidence interval will also depend on the number of people in your sample (more is better), and the SEM (lower is better).

- For evaluation and research purposes, and for high-stakes testing, we set a **90% - 95%** confidence level for the smallest score range with practical implications. In public polling, example, we feels better when we read, "This poll is accurate within plus or minus 3.2 points, 95% of the time." One feel more dubious if we were to read, "This poll is accurate to plus or minus 5.5 points, 85% of the time."
- For moderate- and low-stakes assessments, or any situation where generalization from a sample to a population is not a concern, **80%-90%** confidence levels may be acceptable, and somewhat more generous score ranges may be used.

Validity (Correlation with Other Variables) Coefficients: How High Should They Be?

It depends on the context and variables involved, but higher is always better.

- A correlation between two variables (such as a scale score measuring IPCP attitudes and a test score measuring IPCP teamwork skills) of **0.60** and above would be important and meaningful.
- A correlation less than **0.30** is usually considered to be somewhat to fairly minor.
- The more distant in time and place between a predictor variable (such as attitudes) and its outcome (such as performance in practice), the lower the expected correlation.
- A pattern of correlations between various, similar variables, even if they are in the **0.20 – 0.40** range, might be important and useful if they are in the expected direction.

To understand the magnitude of a validity correlation, square it mathematically. A coefficient of 0.60, squared, is 36 (6 x 6 = 36). That means, 36% of the score variance in your set of data can be "explained" or "accounted for" by the two variables. If the validity coefficient is only 0.30, then only (3 x 3) 9% of the score variance can be explained by these variables – and the rest of the variance is due to other variables, unknown factors, or random error.

Factor Loadings: How High Should They Be?

The higher the loading, the closer the association between that item with the group of items that comprise a common theme or category "factor"). Determining the strength of the factor loadings depends a bit on the number of items and how many factors are identified. One also looks for a pattern where items load most clearly on to only one factor (rather than several).

- Loadings between **0.60** and above are very meaningful, probably indicative of what the cluster of items is really measuring
- Loadings between **0.40 and .60** are considered meaningful
- Loadings of less than **0.30 or 0.40** are generally not considered meaningful.

Sometimes, items will load on to two or more factors. We usually group it with whichever factor loading is higher. If one factor loading is **within 0.05** of another, we consider them equal. (That is, the item could be grouped with either factor.)

Accounting for the Amount of Shared Variance: How High Should it Be?

In a perfect world, all of the variance in a set of data would be attributed to the factors being measured. The amount of test variance that can be attributed to each factor can be reported in terms of percentages (as in "50% of score variance accounted for), or in units called "eigenvalues." The sum of the eigenvalues for all of the factors combined equals the total variance accounted for in the test. In addition to examining variance at the factor level, you can also look at the total amount of variance accounted for by all items combined. As with other statistics, the higher the proportion of variance accounted for, and the higher the eigenvalue, the better. Determining the amount of *unexplained* variance – i.e., variance that cannot be attributed to any of the factors – is as important, however, as parsing out the individual contribution of factors.

- Factors that account for anything over **20%** of total test variance may be making a useful contribution, especially if the total test variance accounted by a combination of factors is high (e.g., 70% or above).
- At the factor level, eigenvalues need to be *at a minimum* **1.0 or higher** to be considered meaningful. Some analysts use 1.5 or even 2.0 as their threshold.

Effect Sizes: How High Should They Be?

Higher is always better. Effect size is influenced by the number of people in the sample (more is better) as well as the raw difference between scores. Based on Cohen (1969):

- Large effects are **0.80** and above
- Moderate effects are between **0.50 and 0.79**
- Small effects are less than **0.50**

Norman and Streiner (2003) lower the bar and qualify small effects as around 0.20.

Item-Total Correlations: How High Should They Be?

The correlation between a single item and the total score can be used to identify items that may not be contributing very much to the internal consistency reliability of the test. Revising or removing them may be helpful for not only improving internal consistency reliability, but shortening the test.

- The same rules of thumb for interpreting the size of any correlation may be used here.
- It is also useful to compare item-total correlations across all of the items in the test, to see which are low compared to the others.
- Most statistical programs will calculate what the internal consistency reliability ("alpha") would be if suspect items are removed. If deleting an item with a "low" (e.g., below 0.60) item-total correlation does not improve the overall reliability of the scale or test, you should probably keep it, unless there are other concerns about the item's relevance or clarity.

Useful References

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Downing, S.M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, 38:1006-12.

Downing, S.M. (2003). Validity: On meaningful interpretation of assessment data. *Medical Education*, 37(9): 830-37.

Norman, G. R & Streinger, D. L. (2003). *PDQ: Pretty Darned Quick Statistics*, 3rd ed. Hamilton / London, BC Decker, Inc.

Vogt, E. P. (2005). *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences*, 3rd ed. Thousand Oaks / London / New Delhi: Sage Publications.