Questions and answers about language testing statistics:

Consistency of measurement categories and subcategories

James Dean Brown
brownj@hawaii.edu
University of Hawai‘i at Mānoa

Question:

This column responds to an email I recently received which raised what is clearly the most concise, even terse, question I have ever received for this column: “Hello....what is the exact difference between external reliability and internal reliability in quantitative research?”

Answer:

I will begin by directly addressing where I think your question is coming from. I will then answer your question by expanding on the notion of consistency. Consistency (or the degree to which something is systematic) is one concern in both measurement and in research design (as shown in Figure 1). In this column, I will discuss consistency in measurement, which comes in two flavors: norm-referenced test (NRT) reliability and criterion-referenced test (CRT) dependability (see the components with the light background). In the next column, I will discuss consistency in research design which comes in three flavors: quantitative reliability, mixed methods research (MMR) dependability, and qualitative dependability (see the components with the darker background).

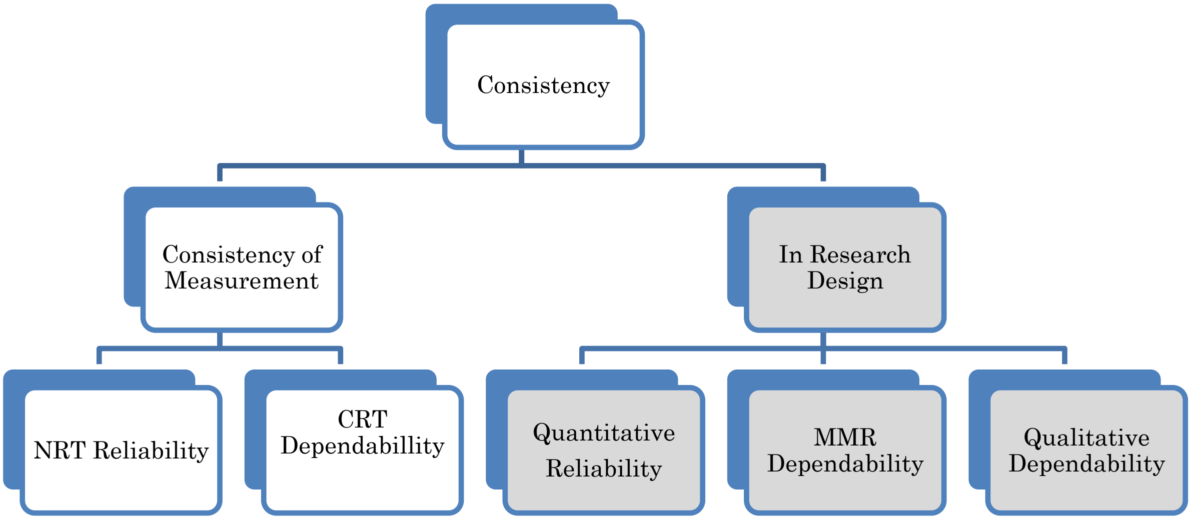

Figure 1. Consistency in measurement and research design

Where I think your question is coming from

In describing *reliability*, some testers/researchers refer to external and internal reliability. *External reliability* is defined by them something like: “the extent to which a measure varies from one use to another” (e.g., test-retest reliability, or interrater reliability); and *internal reliability* is defined something like “the extent to which a measure is consistent within itself" (e.g., split-half reliability) (see McLeod, 2007, no page numbers). Such characterizations of reliability seem to me to be oversimplified and incomplete in terms of ways of categorizing and comparing the various types of consistency that researchers need to cope with in their measurement.

Consistency of measurement categories and substrategies

Here, I will focus on consistency of measurement, which I will define as the degree to which measurements or observations are consistent. I will divide consistency of measurement into the two categories shown to the left in Figure 1: norm-referenced test (NRT) reliability and criterion-referenced test (CRT) dependability. These two will then be further divided into substrategies in order to clarify the different ways there are for looking at consistency within each category.

NRT reliability.

The term *reliability* will only be used in this column to describe the degree of consistency for the sorts of standardized measures that are norm-referenced (for more on this concept, see Brown, 2005) like the TOEFL, IELTS, TOEIC, etc. and are therefore designed to spread people out. The reliability of NRTs can be estimated, corroborated, improved, or verified by using a number of strategies (see Figure 2): *stability over time* as in test-retest reliability or intrarater reliability (e.g., raters at time 1 and time 2); *stability between forms* as in parallel forms reliability (e.g., forms A and B); *stability across scorers* as in interrater reliability (e.g., rater 1 and rater 2); and *stability across items* in a test as in internal consistency reliability (e.g., split-half adjusted, Cronbach alpha, K-R20, K-R21, etc.) (for more on all of these, see Brown, 2005, 2016, pp. 105-138, 149-153; Brown & Hudson, 2002). Naturally, other forms of stability may be of concern. For example, stability across rating categories, rating occasions, tasks, and so forth may be of concern too, but these are really just variations of the four types of stability mentioned in the previous sentence.

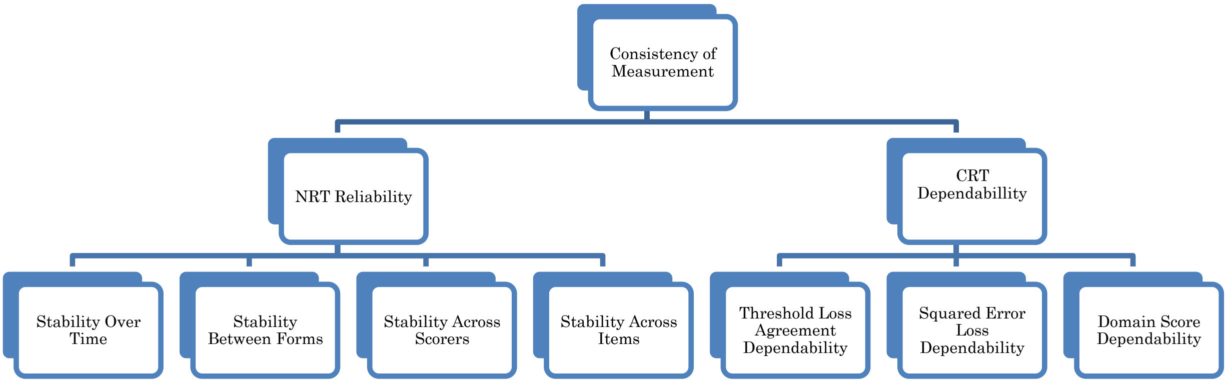
Consistency of Measurement
NRT Reliability
CRT Dependabillity
Stability Over Time
Stability Between Forms
Stability Across Scorers
Stability Across Items
Threshold Loss Agreement Dependability
Squared Error Loss Dependability
Domain Score Dependability

Figure 2. Consistency of measurement: Categories and substrategies

CRT dependability.

In contrast, the term *dependability* will be used here to describe the degree of consistency for measures that are criterion-referenced (for more, see Brown & Hudson, 2002). The dependability of CRTs can be estimated, corroborated, improved, or verified using three strategies (see Figure 2): *threshold loss agreement* dependability (including the agreement and kappa coefficients); squared-error loss agreement dependability (especially the phi lambda coefficient); and *domain-score* dependability (especially the generalizability coefficient also known as the phi coefficient). Note that for the special case of rating or coding language or other samples researchers typically use variations on agreement and kappa coefficients in the threshold loss agreement strategy (see Brown, 2016, pp. 139-147).

Consistency of measurement in research studies

In research studies, the reliability and dependability of measurements and observations can be enhanced by thoughtfully planning, designing, and creating the measures involved. It will also help to pilot and revise any measures before actually implementing them in a research project—all with an eye toward making them reliable or dependable as appropriate. In cases where researchers or their colleagues will be coding or rating data in a study, the reliability/dependability of measures can be enhanced by providing coders/raters with clear guidelines, coding schemes, rubrics, etc., and by providing effective training, or retraining, as may be appropriate.

The place of G theory in consistency of measurement

Those language researchers who learn about language testing analysis typically learn only about classical theory statistics like those discussed above in the section on NRT reliability. Here, I have already pushed beyond that basic knowledge in discussing CRT dependability. However, one step even further away from CTT is the area of Generalizability theory (or G theory, as it is affectionately known). G theory was first proposed by Cronbach and his colleagues at Stanford University (Cronbach, Gleser, Nanda, & Rajaratnam, 1970; Cronbach, Rajaratnam, & Gleser, 1963). G theory has three distinct advantages over CTT. First, it allows for examining multiple sources of error (unreliable variance) in a set of scores. Second, G theory can be used to examine multiple sources of error for either NRTs or CRTs (by using different strategies). Third, G theory can be used to calculate what-if reliability or dependability estimates for different sources of error in terms of numbers of items, raters, occasions, categories, etc. and it can do so for multiple sources of error simultaneously. Thus, G theory supplies an altogether new way of looking at the consistency of scores on any sort of assessment procedures from multiple-choice to task-based. (For more on G-theory, see Brown, 2016, pp. 131-138.)

Conclusion

In direct answer to your question, at least in terms of measurement consistency, "the exact difference between external reliability and internal reliability in quantitative research" is not a very clear, helpful, or adequate way of characterizing the consistency issues of importance.

Here I have shown that consistency in measurement, comes in two forms: NRT reliability (including strategies to study stability across time, between forms, across scores, or across items) and CRT dependability (including threshold-loss agreement, squared error loss agreement, and domain score dependability substrategies). I have also talked about ways to enhance any of those strategies in research studies as well as the place of G theory in this whole framework.

In the next column, I will explain how issues of internal and external validity (and reliability) are important to researchers who want to produce high quality research in our field. To those ends, I will discuss consistency strategies in research design (including quantitative reliability, and mixed methods or qualitative dependability), and how they can be enhanced and corroborated.

References

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment (New Edition)*. New York: McGraw-Hill.

Brown, J. D. (2016). *Statistics corner: Questions and answers about testing statistics.* Tokyo: Testing and Evaluation Special Interest Group of JALT.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge, UK: Cambridge University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1970). *The dependability of behavioral measurements*. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137-163.

McLeod, S. A. (2007). What is reliability? Retrieved from www.simplypsychology.org/reliability.html

Where to submit questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

JD Brown
Department of Second Language Studies University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA