Questions and answers about language testing statistics:

Consistency in research design: 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 that 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:

This is the second of two columns. In both columns, consistency is defined simply as the degree to which something is systematic. I discussed consistency in measurement in the last column as shown in the rectangles to the left with grey backgrounds in Figure 1 in terms of norm-referenced test (NRT) reliability and criterion-referenced test (CRT) dependability. In this column, I will discuss consistency in research design which comes in three flavors: quantitative reliability, qualitative dependability, and mixed methods research (MMR) dependability (see the rectangles to the right with the white backgrounds in Figure 1).

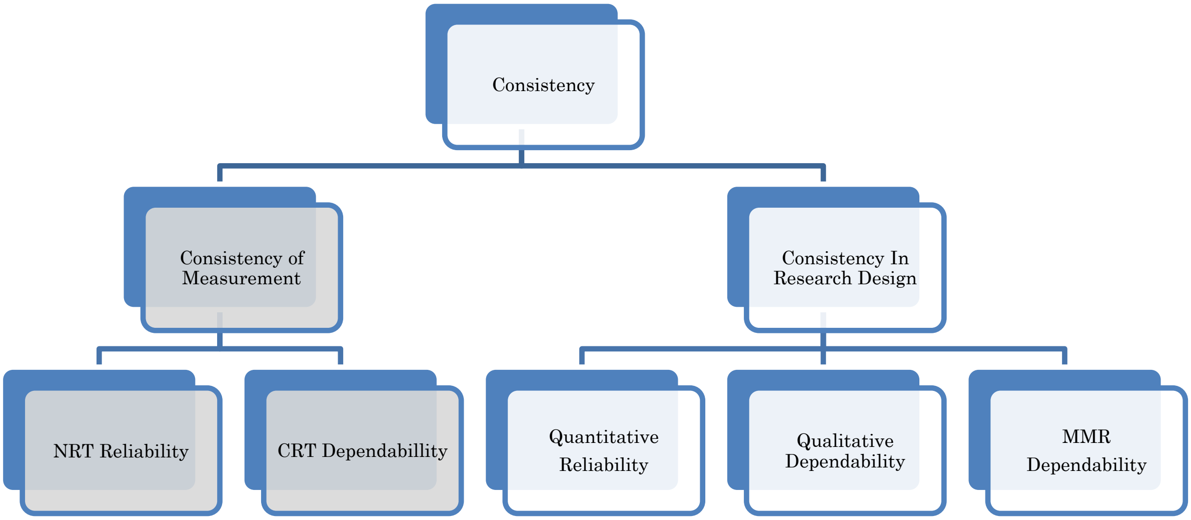

Figure 1. Consistency in measurement (grey) and research design (white)

Consistency in research design categories and substrategies

As mentioned above, consistency in research design falls in three categories: quantitative reliability, qualitative dependability, and MMR dependability, and each of those can be further subdivided into two or three subcategories (as shown in Figure 2).

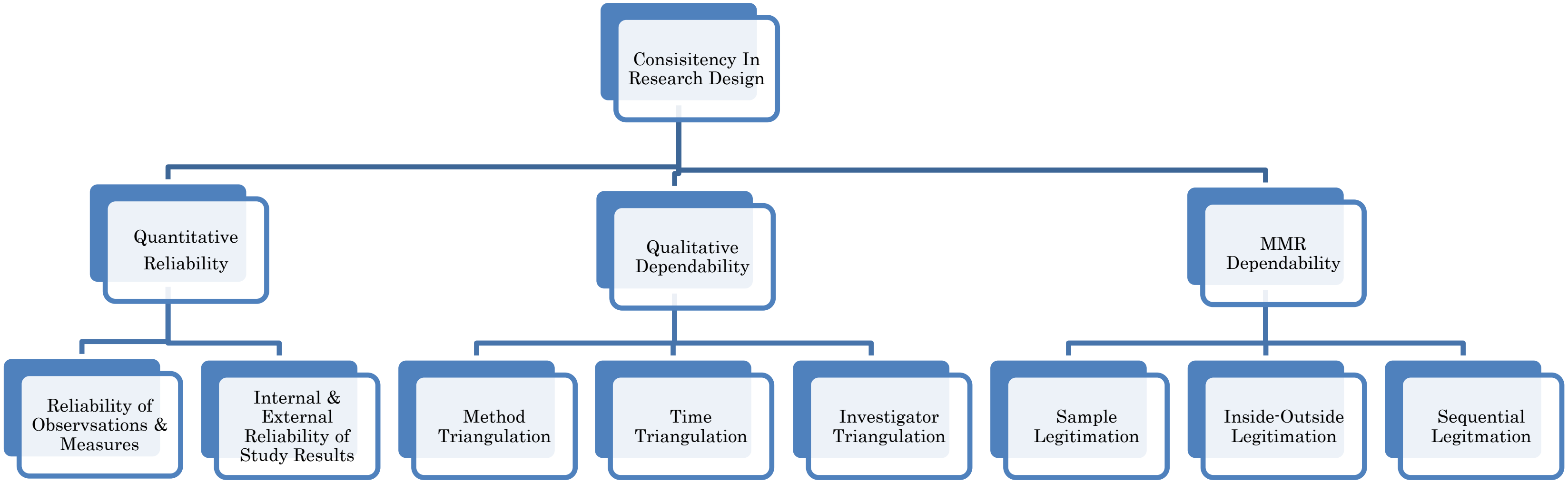
Consisitency In Research Design
Quantitative Reliability
Qualitative Dependability
MMR Dependability
Reliability of Observsations & Measures
Internal & External Reliability of Study Results
Method Triangulation
Time Triangulation
Investigator Triangulation
Sample Legitimation
Inside-Outside Legitimation
Sequential Legitmation

Figure 2. Consistency of in research designs: Categories and subcategories

Quantitative reliability. Quantitative reliability has to do with the degree to which the results of observations/measures are consistent in a study, but also the degree to which the results of the study as a whole are consistent internally and externally. Thus, enhancing or confirming the reliability of a quantitative study should use at least two strategies. The first of these, the reliability of observations/measures strategy can be confirmed or enhanced by calculating reliability estimates for measures or agreement estimates for ratings/codings. For example, for those measures that are based on tests, test–retest, parallel forms, or internal consistency reliability estimates can be calculated, or inter-rater/intra-rater reliability estimates for ratings (for much more on this topic, see Bachman, 2004, pp. 153-91; Brown, 2005, pp. 169-98, 2012), while calculating agreement estimates may be more appropriate for other sorts of observations based on ratings or codings (see Brown, 2001 pp. 231-40).

In contrast, the reliability of the results of the study as a whole can be enhanced *internally* by carefully monitoring and controlling issues that might contribute to inconsistency in design like (a) changes over time due to self-selection of participants into (i.e., using volunteers) or dropping out of the study, (b) maturation in the participants, (c) the Hawthorne effect, (d) the halo effect, or (e) subject/researcher expectancy effects. The reliability of a study can be enhanced/verified *externally* by inspecting the statistical tests that are run in a study with an eye to determining the degree to which the results of the study would be likely to be stable if the study were replicated; for instance, by recognizing that a significant result (at say $p < .01$) means that there is only a 1% chance that the result is due to chance, external reliability can be addressed by thinking about what such probabilities mean for the stability of the results in replication [For definitions and discussion of the terminology used in this paragraph, see Brown, 1988, pp. 29-42, or 2016, pp. 49-53, 162.]

Qualitative dependability. The idea of dependability in qualitative research involves confirming or enhancing the consistency of observations and the effects of changing conditions in the study. Enhancing or confirming the dependability of a qualitative study can use one or more of at least three strategies. The first of these is *method triangulation* (aka overlapping methods), which means using multiple data gathering techniques; for example, a study might include interviews, classroom observations, and a Likert-item questionnaire so that the researcher can examine the dependability of results across methods. The second strategy involves using *time triangulation* (aka stepwise replications), which means gathering data at multiple times; for instance, qualitative data could be gathered at the beginning, middle, and end of a school term so that the dependability of the results over time could be examined. And a third strategy would be to use *investigator triangulation* (aka auditor and inquiry audits), which means having multiple investigators work on the study; for example, qualitative data could be coded by two different investigators with the goal of examining the dependability of codings across investigators. [For more on this terminology and these strategies, see Brown, 2001, pp. 227-231; 2016, p. 158.]

MMR dependability. Since MMR dependability focuses on the consistency of combining quantitative and qualitative data, the consistency of those underlying data and interpretations are a precondition. That is, the reliability of the quantitative data and results should be confirmed or enhanced with regard to the measure/observations and the study as a whole by using the strategies described two subsections above, and the dependability of the qualitative data and results should be confirmed or enhanced with regard to the consistency of observations and effects of changing conditions in the study by using the strategies (i.e., method, time, and investigator triangulation) described in the previous subsection. However, from the additional MMR perspective, the dependability of the efforts to combine quantitative and qualitative data should be examined using at least three types of legitimation: sample, inside-outside, and sequential legitimation. *Sample legitimation* involves examining or enhancing the ways that the qualitative and quantitative samples were integrated and consistent within a study; for example, by examining the consistency of results from qualitative interviews and classroom observations, then examining the quantitative Likert item questionnaires developed from those interviews and observations, and checking all of that with qualitative follow-up interviews used for member checking. *Inside–outside legitimation* involves considering how adequately the insider (emic) and outsider (etic) perspectives were combined in the quantitative and qualitative data and analyses; for instance, by studying the degree to which the emic perceptions of students and teachers in an institution gathered in qualitative interviews compared or combined with the etic perceptions of the public about that institution gathered in quantitative Likert item questionnaires. *Sequential legitimation* examines the degree to which the effects of method sequencing were minimized; for example, by considering the degree to which results based on interviews conducted before and after the administration of the Likert item questionnaire were consistent. [For more on the concepts discussed in this paragraph, see Brown, 2014, especially pp. 127-135.]

Table 1

Summary of Research Consistency Categories and Subcategories in Quantitative, Qualitative, and MMR Research with Examples

Type	Category	Subcategory	Example
Quantitative Reliability	Reliability of Observations & Measures	Enhanced/confirmed by calculating reliability estimates for measures; or calculating agreement coefficients for ratings or codings	For tests, calculating reliability estimates like test-retest, parallel forms, or internal consistency (e.g., Cronbach alpha, K-R20, etc.) or inter-, or intra-rater reliability estimates; For other sorts of observations, calculating rater/coder agreement coefficients or kappa
	Internal & External Reliability of Study Results	Internal reliability– enhanced/confirmed by controlling issues that often contribute inconsistencies in study design	Monitoring & controlling issues like self-selection, mortality, maturation, Hawthorne effect, halo effect, or subject/researcher expectancies
		External reliability – enhanced/verified by inspecting statistical results in terms of replication	Recognizing that a significant result (at say $p < .01$) means that there is only a 1% chance that the result is due to chance, external reliability can be addressed by thinking about what such probabilities mean for the stability of the results in replication
Qualitative Dependability	Method Triangulation	(aka overlapping methods) Enhanced/confirmed by using multiple data gathering methods	For example, using interviews, classroom observations, & a Likert item questionnaire, & examining dependability of results across methods
	Time Triangulation	(aka stepwise replications) Enhanced/confirmed by gathering data at multiple times	For example, gathering data at beginning, middle, & end of school term, & examining dependability of results over time
	Investigator Triangulation	(aka auditor & inquiry audits) Enhanced/confirmed by using multiple investigators	For example, using two investigators to independently code the data in a study, & examining dependability of results across investigators
MMR Dependability	Sample Legitimation	Enhanced/confirmed by examining how the qualitative & quantitative data samples are integrated & consistent	For example, examining the consistency of results from qualitative interviews & classroom observations, quantitative Likert item questionnaires developed from those interviews & observations, & qualitative interviews used for member checking later in the study
	Inside-Outside Legitimation	Enhanced/confirmed by considering how adequately the insider (emic) & outsider (etic) perspectives were combined in the quantitative & qualitative data & analyses	For instance, by studying the degree to which the emic perceptions of students & teachers in an institution gathered in qualitative interviews compared or combined with the etic perceptions of the public about that institution gathered in quantitative Likert item questionnaires
	Sequential Legitimation	Enhanced/confirmed by examining the degree to which the effects of method sequencing were minimized	For example, by considering the degree to which results based on interviews conducted before & after the administration of the Likert item questionnaire were consistent

Conclusion

In direct answer to your question, “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 that arise in of consistency of measurement or consistency of research design.

In the previous column, I addressed the issues involved in consistency of measurement separately for norm-referenced and criterion-referenced tests. In the present column, I have shown that consistency in research design, comes in three categories: quantitative reliability (with subcategories for consistency of observations and measurement and consistency of study results internally and externally), qualitative dependability (with subcategories for method, time, and investigator triangulation), and MMR dependability (with subcategories for sample, inside-outside, and sequential legitimation). Table 1 summarizes all of those aspects of consistency in research.

I hope this and the preceding column together have addressed your question and helped you to realize that a simple external/internal reliability categorization of the issues involved is neither complete nor useful.

References

Bachman, L. F. (2004). *Statistical analysis for language assessment.* Cambridge: Cambridge University.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher’s guide to statistics and research design.* Cambridge: Cambridge University.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

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

Brown, J. D. (2014). *Mixed methods research for TESOL.* Edinburgh, UK: Edinburgh University.

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

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