Questions and answers about language testing statistics:

Characteristics of sound quantitative research

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

QUESTION:

In Brown, 2005, you explained the characteristics of well-done qualitative research by explaining the importance of dependability, credibility, confirmability, and transferability. You mentioned in passing that the parallel characteristics for quantitative research were reliability, validity, replicability, and generalizability. But you never really explained those quantitative research characteristics. I think it would be useful to know more about those characteristics of sound quantitative research and maybe even something about the characteristics of good quality mixed-methods research. Could you talk about these other research paradigms?

ANSWER:

Certainly, let me begin by reviewing my definition of what I think research is. Then I will turn to the issues that quantitative researchers need to address in order to produce sound quantitative research by explaining four concepts: reliability, validity, replicability, and generalizability. As I proceed through these explanations, you will see how similar and yet different the qualitative and quantitative sets of characteristics are. I will focus on the characteristics of quantitative research here and save the characteristics of mixed-methods research for a subsequent column (Brown, forthcoming in 2016).

What is research?

In the column you refer to (Brown, 2005), I defined research very broadly as: "any systematic and principled inquiry" (based on Brown, 1992, 2004). Research can be systematic and principled in many different ways. As I discussed in Brown (2005), sound qualitative research (at one end of the continuum) can be systematic in terms of its dependability, credibility, confirmability, and transferability, while sound quantitative research can be systematic in terms of its reliability, validity, replicability, and generalizability—four characteristics that will serve as the focus of the rest of this column.

Reliability

In quantitative research, at a micro level, reliability can be defined something like the degree to which the results of research measurements and observations are consistent. The reliability of a study's measurements and observations can be enhanced by carefully designing and creating them, piloting them beforehand, and revising them with an eye toward increasing their reliability before they are ever used in the main study. In cases, where humans will be rating or coding data, reliability may be enhanced by giving the raters/coders clear guidelines, carefully training them, and periodically retraining them (especially if the ratings will be done over a long period of time).

The reliability of a study's measurements and observations can be checked in cases were test items or Likert-item questionnaires are involved, either by calculating test-retest reliability (i.e., examining the degree of correlation between the scores produced by two administrations of the same test or questionnaire), parallel forms reliability (i.e., examining the degree of correlation between the scores

produced by two forms of the same test or questionnaire), or more easily, by calculating internal consistency reliability estimates (e.g., Cronbach alpha, K-R20, etc.) as appropriate. Alternatively, in cases where the measurements or observations are being assigned by raters, interrater reliability can be used (typically by examining the degree of correlation between ratings assigned by pairs of raters), and when the measures or observations are being coded by human coders, intercoder agreement will be used (typically, by calculating the percent of codings that agree between two coders).

However, at a macro level, reliability can also be defined as the degree to which the results of a study are consistent. This type of macro reliability can be enhanced by carefully (a) sampling, (b) thoughtfully planning and controlling the conditions under which the study is conducted, and (c) meticulously designing, piloting, and revising all measurement and observation tools. In general, then, the reliability of a study should be examined in terms of how well the results of the study are internally consistent and make sense in terms of sampling, study conditions, and instrumentation.

Validity

In quantitative research, at a micro level, validity can be defined as the degree to which a study's measurements and observations represent what they are supposed to characterize. The validity of a study's measurements and observations can be enhanced by carefully designing and creating them based on the best available language learning theories, piloting them beforehand, and revising them with an eye toward increasing their validity in terms of how accurately they are measuring what they were intended to measure.

The validity of the scores or other values obtained from any instrumentation in a study can be checked and/or defended by studying evidence and developing arguments for the content, criterion-related, construct validity of the resulting scores or other values, as well as their social consequences and values implications within the study and more broadly.

At a macro level, validity can also be defined as the degree to which the results of a study represent what the researcher thinks they represent. This type of macro validity can be enhanced by initially designing a study to maximally approximate "natural" conditions; by carefully prearranging and controlling study conditions; and by guarding against effects like the Hawthorne effect, halo effect, subject expectancy effect, researcher expectancy effect, practice effect, and reactivity effect (see Brown, 1988, or many other sources).

Replicability

Replicability can be defined as the degree to which a study supplies sufficient information for the reader to verify the results by replicating or repeating the study. The replicability of a study can be enhanced by writing a clear and complete research report in the style of a recipe that tells readers about: the participants (including who they were and how they were selected), the materials (including what measurements and observations were used in the study and why they were reliable and valid for that purpose), the procedures (including all of the steps in how the study was conducted), and the analyses (including how the variables were defined and arranged, as well as all analyses that were performed to address the research questions). Indeed, the study should be so clearly described that a reader could in fact repeat the study if they were so inclined. One way to check this is to ask a colleague to read the report and give you feedback with the notion of replicability (as described here) in mind.

Generalizability

Generalizability can be defined as the degree to which the results of a study can be generalized, or are meaningful, beyond the sample in a study to the population that the sample represents. Unfortunately, it

is often very difficult to define a general population in second language studies. For example, in an ESL study, can we ever say that a sample of students selected from the English Language Institute at the University of Hawai'i at Mānoa (UHM) is representative of all ESL students studying in the US? Or even all ESL university students studying in the Hawai'i? Can we say that this predominantly Asian sample of international students is the same or even similar to ESL students studying at a US East Coast university, where students might tend to be predominantly European and Middle Eastern? I think you can see the problem.

However, there is no reason to lose hope because the generalizability of a study can be enhanced in at least four ways:

- Narrowly define the population you are trying to sample from. For example, don't even pretend that you are trying to generalize to all ESL students in US universities (or even to all EFL students in Japanese universities). Instead, define the population narrowly as in the population of all students in the ELI at UHM. Then and only then will it be reasonable to say that a sample selected randomly or in a stratified manner represents that population of students in the ELI at UHM.
- Choose participants with random or stratified selection into the study and then into whatever groups you may want to compare (e.g., treatment and control groups). Those strategies will definitely help to improve the representativeness of the sample(s) and thus the generalizability of the study (see Brown, 2006).
- Control for self-selection and mortality of participants (a) by avoiding the use of volunteers whenever possible (i.e., self-selection) and (b) by minimizing as much as possible all attrition (i.e., participants dropping out of the study, also known as, mortality) by keeping the study short in duration and by encouraging participants to stay in the study. The reasoning here is that people who volunteer tend to be a certain type of gung-ho student not representative of the entire population, and similarly, people who leave a study or drop out may also be a certain type of person who by leaving will make the remaining participants less representative.
- Use the qualitative concept of transferability described in Brown (2005), which was described as follows: "Transferability can be enhanced by providing what is often referred to as thick description (i.e., giving enough detail so the readers can decide for themselves if the results are transferable to their own contexts)" (p. 32). What I am saying is that providing readers with very clear information about who the participants were and how they were selected will help those readers determine for themselves how much the results can be generalized, or better yet, how much the results may apply to their own teaching/research situations.

Conclusion

In direct answer to your original question, the characteristics of sound quantitative research are generally considered to be: reliability, validity, replicability, and generalizability.

These are of course ideals that researchers should strive for and of course may be enhanced, or defended in a variety of different ways depending on the type of study, the research questions involved, the nature of the variables, the choices of statistical analysis techniques, and so forth. Because these characteristics are ideals, they can also serve as standards against which you as a reader can judge the quality of quantitative research that you encounter in our ever growing literature. And of course, remember to apply these same standards just as critically to any research that you yourself may produce.

Those readers who find quantitative research methods intriguing may find it useful to read books like Baayen (2008), Brown (1988, 2001), Brown and Coombe (2015), Brown and Rodgers (2002), Butler (1985), Dörnyei (2003, 2007), Hatch and Lazaraton (1991), Porte (2010), Rietveld and van Hout (1993),

and Scholfield (1995); and, those interested in moving beyond the basic level should consider reading Plonsky (2015) and perhaps even Tabachnick and Fidell (2012).

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University.

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

Brown, J. D. (1992). What is research? *TESOL Matters*, *2*(5), 10.

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

Brown, J. D. (2004). Research methods for applied linguistics: Scope, characteristics, and standards. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 476-500). Oxford: Blackwell.

Brown, J. D. (2005). Statistics Corner. Questions and answers about language testing statistics: Characteristics of sound qualitative research. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *9*(2), 31-33. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_22.htm

Brown, J. D. (2006). Statistics Corner. Questions and answers about language testing statistics: Generalizability from second language research samples. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *10*(2), 24-27. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_24.htm

Brown, J. D. (forthcoming in 2016). Characteristics of mixed methods research. *Shiken Research Bulletin*, *20*(1).

Brown, J. D., & Coombe, C. (Eds.) (2015). *The Cambridge guide to research in language teaching and learning*. Cambridge: Cambridge University.

Brown, J. D., & Rodgers, T. (2002). *Doing second language research*. Oxford: Oxford University.

Butler, C. (1985). *Statistics in linguistics*. Oxford: Blackwell.

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Rowley, MA: Newbury House.

Plonsky, L. (Ed.) (2015). *Advancing quantitative methods in second language* research. New York: Routledge.

Porte, G. K. (2010). *Appraising research in second language learning: A practical guide to critical analysis of quantitative research*. Amsterdam: Benjamins.

Rietveld, T., & van Hout, R. (1993). *Statistical techniques for the study of language and language behaviour*. Berlin: Mouton de Gruyter.

Scholfield, P. (1995). *Quantifying language: A researcher's guide to gathering language data and reducing it to figures*. Clevedon, UK: Multilingual Matters.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). New York: Pearson.

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