EDITORIAL

Approaches for Improving Validity in Quantitative Research Articles

Andrew D. Asher

Introduction

Validity, or the assessed likelihood that a research design represents and measures the concepts it purports to study, is an essential component of evaluating the quality and efficacy of research manuscripts prior to publication. Since there is no single method to determine validity, readers must make a comprehensive judgment based on both the conceptual and technical elements of a research design, making it imperative that authors provide sufficient contextualization and discussion of data collection and analysis decisions that might affect the interpretation of their findings and conclusions.

Especially when using well-established statistical procedures, manuscripts presenting quantitative analysis often gloss over information that is fundamental to understanding the underlying validity of the research. A three-step evaluation heuristic that considers the sample, the statistic and significance, and the effect size is a useful tool to help correct for this problem and to ensure that key elements are present, appropriate for the research design, and contain sufficient information to assess validity (See Figure 1).

portal: Libraries and the Academy, Vol. 24, No. 2 (2024), pp. 209–215

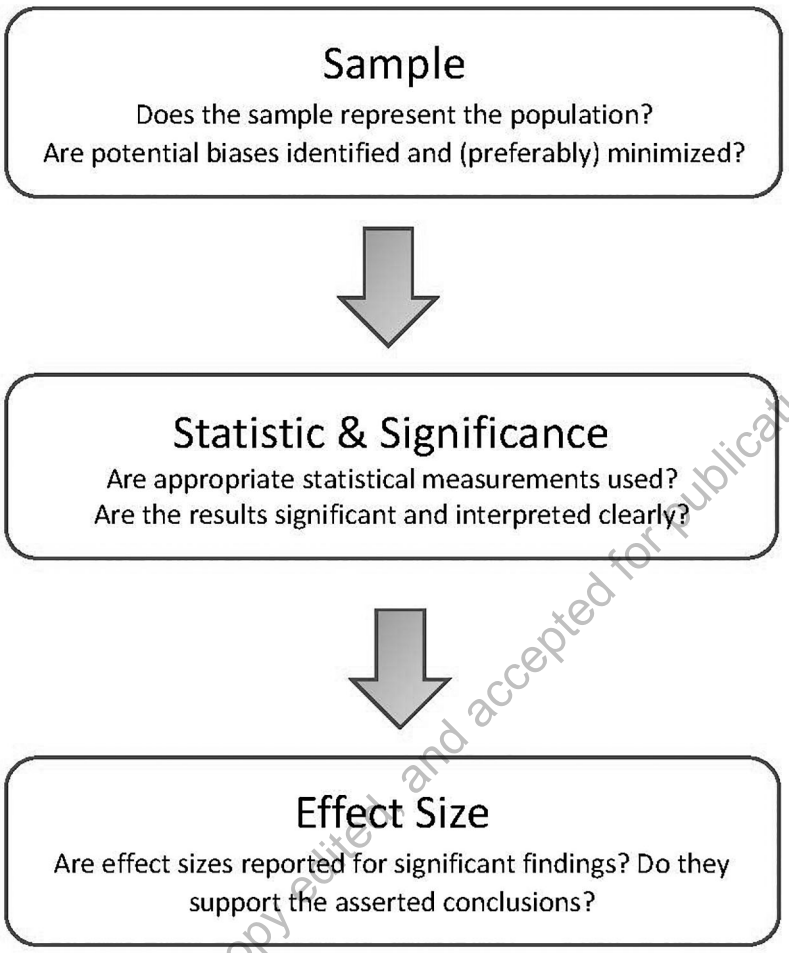

Figure 1. A "4S" evaluation workflow for quantitative analysis: Sample, Statistic & Significance, and (Effect) Size.

Using this heuristic as a starting point, this editorial discusses common problems found in quantitatively focused manuscripts with the goal of providing guidance to prospective *portal* authors prior to review. Many of these errors and omissions stem from the types of quantitative research designs that are frequently employed in LIS research, such as methods involving surveys or rubrics, but the principles discussed here are broadly applicable to other research approaches as well.

Sampling Design: A Study's Foundation

Draft manuscripts often include insufficiently described sampling methods. The quality of sampling procedures and how well a sample represents the population under study provide the foundation for the validity of all subsequent inferences based on the data collected. Clear and detailed discussion of the sampling approach and its limitations is

therefore necessary as a first step to enable reviewers and readers to evaluate the quality of the research presented.

Probability-based sample designs are usually required to ensure that analyses involving inferential statistics are valid, and so that concepts such as uncertainty, confidence, and risk can be quantitatively represented. Probability-based sampling methods range from simple to complex, but in all cases the chance of a member of the target population being included in the sample is known, can be expressed as a probability, and is a result of a randomization procedure.[1] However, while probability-based design is almost always a best practice, practical constraints such as cost, time, or data availability often make these designs difficult to implement, resulting in researchers relying on non-random or inadequately randomized samples.

Survey-based methods can be particularly vulnerable to this issue, and a frequent problem in LIS studies is the presentation of survey results as if they are based on probability samples when they are not. Surveys sent to listservs are often an example of this error. LIS surveys routinely use listservs as a proxy for the population of library professionals in place of randomized sample designs. This practice results in several potential problems. First, without detailed information about the membership of a listserv and associated demographic data, it is not possible to accurately estimate how well it represents the population under study. This produces a high potential for selection bias—members of the listserv may differ in some systematic way from the population at large—as well as non-response bias—people who answer surveys on listservs may differ from people who do not.[2] Second, since the overall membership of a listserv is often unknown to the researchers, it is impossible to calculate common measures of survey quality such as response rates and margin of error, which are essential to evaluating both validity and confidence in the survey outcomes. At best, responses from a survey sent to a listserv can only represent the population of the listserv itself—without additional information we cannot interpolate how well any group of respondents potentially represents a wider population.

Utilizing a listserv for gathering survey responses is an example of a type of non-probability or convenience sampling. A strict methodologist might argue that, formally speaking, no inferential statistics are valid in a quantitative study based on such a design. Furthermore, the associated risk of an incorrect inference cannot be known using a non-probability sample and might be much higher than the typical maximum five percent threshold (a p-value < .05).[3]

Nevertheless, such designs are very common in the LIS literature, not only for survey research, but also in other quantitative studies because of the real-world constraints many researchers face. For this reason, a non-probability sampling design is not disqualifying in a quantitatively oriented manuscript submitted for review. However, authors should seek to address the problems inherent in these designs by attempting to maximize diversity and variation in non-random sampling through approaches such as purposive and quota sampling that evaluate a collected sample against known population information or demographic data collected during a study. Authors should also include a thorough discussion of a sample's potential limitations, biases, and effect on any statistics presented, as well as any available documentation of how adequately the sampled population (for example, a listserv) reflects the population under study.

Finally, researchers should be careful to note that statistical inferences based on non-probability samples are technically approximations that might be better described as "indications" rather than "estimates" so that the uncertainty contained in this approach is made clear to readers.[4]

Underlying Statistical Assumptions

As with validity questions raised by incomplete examination of non-probability samples, many quantitative manuscripts do not provide enough evidence that the authors have considered how the underlying assumptions of statistics they choose might affect the validity of their analysis. This is particularly true in manuscripts presenting results from variables measured at the ordinal level.

Many LIS studies use Likert or comparably designed scales as a means of measuring variables such as opinions, agreement with statements, attitudes, or behaviors. These scales, as well as many other measurements applied to educational outcomes—such as rubric evaluation scores or letter grades—are measured at the ordinal level, that is, the scales are ordered, but the distance between the points is not theoretically uniform. For example, the distance between "satisfied" and "very satisfied," and "very satisfied" and "extremely satisfied" are not necessarily conceptually equal, nor is an "extremely satisfied" person experiencing twice the satisfaction of a "satisfied" person. In contrast, interval level measurements are characterized by scale points that are at a uniform distance from one another (millimeters on a ruler, for example). This important difference often becomes muddled when ordinal scales are recoded as integer values during analysis.

Strictly speaking, no arithmetic or computation based on ordinal measurements is conceptually valid, including measures routinely reported for ordinal scales such as mean and standard deviation, nor are any inferential statistics that assume parametric data. Parametric statistics assume that the distribution of the data under study is normally distributed with defined parameters based on the shape of the distribution (for example, 95 percent of the data is within two standard deviations from the mean) and that the data is measured at the interval level. Study designs using ordinal scales typically violate both assumptions, making their data most appropriately analyzed using non-parametric statistics that do not assume a normal distribution (See Table 1).

However, in LIS, education, and other social science literature, ordinal scales are routinely treated as interval measurements during analysis, and there is extensive debate about the conditions under which ordinal data can be handled in this way.[5] Geoff Norman reviews the literature applying several parametric procedures to non-parametric data and argues that in most cases they are sufficiently robust that the chance of drawing an incorrect conclusion is minimal even when their assumptions are violated.[6] Composite variables combining multiple ordinal-scale measurements are also less vulnerable to violations of parametric assumptions and can typically be treated as interval measures (as in summing the ratings of a series of questions designed to measure a common construct, assuming appropriate levels of reliability are achieved—this is essentially the process of calculating a GPA based on a group of ordinally measured course grades.)[7] Increasing the number of scale points also appears to make ordinal scales behave in a more parametric manner. Eren Can Aybek and Cetin Toraman illustrate that five- and

Table 1.

Example parametric and non-parametric statistics for common analysis approaches

Analysis	Parametric Statistic Example	Non-parametric Statistic Example
Comparing 2 Dependent Samples	Paired t-test	Wilcoxon signed-rank test
Comparing 2 Independent Samples	Unpaired t-test	Mann-Whitney U test
Comparing 3 or more Independent Samples	ANOVA	Kruskal-Wallis test
Correlations between variable	Pearson's R	Spearman's Rank Correlation

Note: There are many statistics to choose from depending on the characteristics of the data analyzed. Authors should verify which statistic is most appropriate for their design.

seven-point scales perform better than three-point scales, while Shing-On Leung and Huipin Wu and Shing-On Leung show that 11point scales meet conditions for normality but do not perform substantially better than scales with fewer points in other measures of validity.[8]

In short, decisions about scale design and measurement can have substantial and complex effects on the subsequent statistics based on them which should be carefully considered. Increasing the number of scale points appears to decrease risks associated with using parametric statistics when non-parametric statistics are formally appropriate.

Potential measurement problems should be addressed at both the design and analysis phase of a research study. Researchers must balance their quantitative analysis requirements with the realities of data collection in practices. While increasing the number of scale points or creating composite measures might be appropriate in some designs, a 7- or 11-point evaluation rubric is probably neither desirable nor effective (and would likely create other problems such as difficulties in obtaining sufficient interrater reliability). Authors should therefore carefully explain why a particular analysis approach is appropriate for the way data is measured in their design, especially when a decision has been made to utilize parametric statistics when the underlying data makes non-parametric statistics more formally appropriate. This is particularly necessary when drawing conclusions about differences between groups based on averages, distributions, or correlations using data collected via ordinal scales since this is where the risk of error is most acute.

Practical Significance—the Importance of Effect Size

When using a null hypothesis significance testing framework to make inferences about research data, effect size measures are one of the most important statistics for interpreting the likely real-world meaning of observations that are found to be statistically significant. Unfortunately, these measures are also one of the most often omitted from manuscripts presenting inferential statistics. For example, in a review of five years of LIS literature, Juris Dilevko found only two examples of effect size in 69 articles presenting statistical results.[9]

In the null hypothesis testing approach, the significance value (p) measures the likelihood that the result of a statistical procedure is due to random chance. Because they state probability, p-values have no inherent size attached to them and make no assertion about the magnitude of observed differences. Therefore, when drawing conclusions about observed relationships or differences between variables, effect size is critical to understanding whether these relationships and differences exist in a practical sense or are likely an artifact of statistical noise.

Statistical significance is also related to sample size. The larger a sample, the smaller the observed difference between groups required to produce a statistically significant result, and the more spurious correlations will occur between variables.[10] Calculating and reporting effect size statistics are one check on these potential errors and are especially important when comparing means between groups, such as reporting the Cohen's *d* effect size measure alongside a t-test (The appropriate effect size measure to report will depend on the statistical method used).

Large effect sizes support an interpretation that a meaningful difference exists between groups, or a strong association or interaction exists between variables under study, while a small effect size suggests that no such relationships exist or are very weak, even if they are statistically significant. A larger effect size also diminishes the chance of type I error—observing a relationship when none exists—error that is especially important to avoid when considering high impact interventions or claims about results. Because of its importance to interpreting the validity of statistically supported conclusions, authors using inferential statistics should always include and discuss an effect size calculation.

Conclusion

Designing research for validity is a holistic process that must be considered at all stages of a research study. Since decisions in earlier phases of data collection affect the validity of all subsequent findings, researchers should plan their methods for measurement, sampling, and statistical analysis before any data is collected and examine what biases, limitations, or uncertainty their design decisions may introduce. Authors should be diligent and transparent about including a discussion of these decisions in their manuscripts so that readers and reviewers can understand and assess the validity of assertions and conclusions.

This editorial has outlined some frequent errors and omissions in quantitative LIS articles with the goal of providing a framework for assisting authors in their research design and presentation practices. Nevertheless, since real-world research can be messy

and contingent on outside constraints, it is difficult to provide recommendations that are applicable to every research study; trade-offs are often a necessity when weighing the resources allocated to a project with the level of uncertainty or risk of incorrect conclusion that can be accepted in a study. The judgment of the researcher is central in these considerations and the details of their thought process should be reflected in their manuscripts. Samples, statistics, and effect sizes are three areas where it is essential for authors to provide sufficient information and discussion to outline the extent of validity in their research design.

Andrew D. Asher (he/him) is the Assessment Librarian at Indiana University, Bloomington and a member of the portal editorial board. He can be reached at asherand@iu.edu. His ORCID is 0000-0002-8600-2191.

Notes

1. For an extensive discussion of sampling design see Yves Tille and Alina Matel, "Basics of Sampling for Survey Research," in *The Sage Handbook of Survey Methodolofy*, ed. Christof Wolf et al. (Los Angeles: Sage Reference, 2016), 311-328.
2. Ronald D. Fricker, "Sampling Methods for Web and E-mail Surveys." In *The SAGE Handbook of Online Research Methods*, eds. Nigel Fielding, Raymond M. Lee, and Grant Blank, (London: SAGE, 2017), 195-217; Jane Fielding and Nigel Fielding, "Synergy and Synthesis: Integrating Qualitative and Quantitative Data," in *The SAGE Handbook of Social Research Methods*, by Julia Brannen, Pertti Alasuutari, and Leonard Bickman (SAGE, 2008), 555–71, https://doi.org/10.4135/9781446212165.
3. See Vasja Vehovar, Vera Toepoel, and Stephanie Steinmetz, "Non-Probability Sampling," in *The Sage Handbook of Survey Methodology*, ed. Christof Wolf et al. (Los Angeles: Sage Reference, 2016), 329–45.
4. Ibid., 334. They also suggest that describing inferences based on non-probability samples "indications" rather than "estimates" might further clarify this issue.
5. Geoff Norman, "Likert Scales, Levels of Measurement and the 'Laws' of Statistics," *Advances in Health Sciences Education* 15, 5 (2010): 625–32, https://doi.org/10.1007/s10459-010-9222-y; Thomas R. Knapp, "Treating Ordinal Scales as Interval Scales: An Attempt to Resolve the Controversy," *Nursing Research* 39, 2 (1990): 121–23; Godfrey Pell, "Use and Misuse of Likert Scales," *Medical Education* 39, 9 (2005): 970–970, https://doi.org/10.1111/j.1365-2929.2005.02237.x.
6. Norman, "Likert Scales, Levels of Measurement and the 'Laws' of Statistics."
7. James Carifio and Rocco Perla, "Resolving the 50-Year Debate around Using and Misusing Likert Scales," *Medical Education* 42, 12 (2008): 1150–52, https://doi.org/10.1111/j.1365-2923.2008.03172.x.
8. Eren Can Aybek and Cetin Toraman, "How Many Response Categories Are Sufficient for Likert Type Scales? An Empirical Study Based on the Item Response Theory," *International Journal of Assessment Tools in Education*9, 2 (2022): 534–47, https://doi.org/10.21449/ijate.1132931; Shing-On Leung, "A Comparison of Psychometric Properties and Normality in 4-, 5-, 6-, and 11-Point Likert Scales," *Journal of Social Service Research* 37, 4 (2011): 412–21, https://doi.org/10.1080/01488376.2011.580697; Huiping Wu and Shing-On Leung, "Can Likert Scales Be Treated as Interval Scales?—A Simulation Study," *Journal of Social Service Research* 43, 4 (2017): 527–32, https://doi.org/10.1080/01488376.2017.1329775.
9. Juris Dilevko, "Inferential Statistics and Librarianship," *Library & Information Science Research* 29, 2 (2007): 209–29, https://doi.org/10.1016/j.lisr.2007.04.003.
10. Cristian S. Calude and Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* 22 (2017): 600–602.