Published in final edited form as:

Nurs Res. 2015 ; 64(2): 137–139. doi:10.1097/NNR.0000000000000078.

“Conceptions of Reliability Revisited, and Practical Recommendations”

Byron Gajewski, PhD [Professor of Biostatistics],

University of Kansas Medical Center Department of Biostatistics Kansas City, KS

Larry R. Price, PhD [Professor of Psychometrics & Statistics], and

College of Education, Department of CLAS College of Science, Department of Mathematics
Director Interdisciplinary Initiative for Research Design & Analysis Texas State University San
Marcos, TX

Marjorie Bott, RN, PhD [Associate Professor and Associate Dean for Research]

University of Kansas School of Nursing Kansas City, KS

Abstract

Sijtsma and van der Ark (2015) present a broad set of models and methods for reliability estimation and their discussion of similarities and differences provides clear information for nurse researchers to move forward in their instrument development projects. In particular, we applaud the authors' clear exposition of the factor analytic model and its utility for providing a framework for unifying reliability and validity. However, we do not want to be constrained only to the point estimates. We also need to ascertain the uncertainty in the point estimate—usually in the form of a 95% confidence interval—or as the Bayesians refer to a credible interval. Another issue not discussed by Sijtsma and van der Ark is conditional standard errors of measurement (CSEM) along the score scale measuring latent traits or true scores. In our response, practical tools for estimating intervals and a brief discussion of CSEMs are presented.

Keywords

Bayes; factor analysis; instrument development; intervals; measurement; reliability

Congratulations to Klaas Sijtsma and Andries van der Ark (2015) for a clearly articulated exposition of measurement or score reliability. Their exposition seamlessly links theory and methods of estimation of score reliability based on the three primary psychometric approaches used in the behavioral (and health) sciences. Reliability is extremely important in many areas of health research. Applications in nursing research, for example, include the estimation of the “score” or “value” reliability using instruments designed to measure the patient's or caregiver's experience under various treatment and/or care conditions. Estimates of reliability support the dissemination and use of new instruments in nursing research (Gajewski et al., 2014; Jiang et al., 2014) by providing one piece of evidence of the

psychometric adequacy of an instrument (American Education Research Association [AERA], American Psychological Association [APA], National Association of Measurement in Education [NCME], 2014). Often asked by those who design studies to estimate score reliability of an instrument is *what is the margin of error of such an estimate?*

A powerful approach to estimate the reliability of scores produced by an instrument is the use of different models. Such models can, for example, lead to a better understanding of the underlying structure of the instrument (e.g., examining content-construct congruence through item homogeneity analysis), and to give a sense of the level of confidence in the reliability estimates.

Investigators explore confidence in a study's reliability estimate by calculating a 95% confidence interval of the reliability estimate. In the traditional models for estimating score reliability, confidence intervals are derived based on classical (i.e., frequentist) probability theory. Bayesian probability offers an alternative to the classical probability school in estimating the amount of error around an estimate of score reliability. Bayesians have probability distributions of reliability estimates and calculate a 95% credible interval—which represents a summary of the reliability parameter. Not surprisingly, a Bayesian that uses no prior information will have an interval that is very close to the non-Bayesian interval. Our own experience is that studies that follow instrument development guidelines always have prior information available in the form of expert opinion that they elicit during the assessment of an item's relevancy in measuring a domain (i.e., content validity). Such information is valid—based on empirical evidence (Gajewski et al., 2012)—when the content experts are approximately on target with assessing the relevance of the items, and then translating that information to a “prior” for information about an item's contribution to score reliability.

With the acceleration of modern computing—regardless of the model that is used for estimating score reliability—deriving an interval for such estimates not only is becoming routine, but is a necessary step to understand where we are in the march toward recommending that an instrument is ready for dissemination. For example, is the amount of error in our reliability estimate too large to make scores from the instrument useful? Often lost in the shuffle of what model to use for this process is the important step in calculating whether we have enough confidence to make such a call.

Sijtsma and van der Ark present a broad set of models and methods for reliability estimation, and their discussion of similarities and differences provide clear information for nurse researchers to move forward in their instrument development projects. For example, their comment focuses on the perspectives of: (a) classical true score theory (CTT); (b) factor analysis; and (c) generalizability theory (Brennan, 1983; 2010); and touches on item response theory (IRT; de Ayala, 2009) approaches. But while this exposition provides closer movement toward connecting principles—even providing examples with data using the free and valid statistical software R—there are several noteworthy issues missing, and we address these in this response. For example, Sijtsma and van der Ark present three primary psychometric techniques for estimating *point estimates* of score reliability. However, the authors do not provide direct advice on how an investigator might calculate 95% confidence

intervals from these models. Additionally, their presentation does not address conditional standard errors of measurement (CSEMs) along the score scale. Conditional standard errors of measurement (and person- or ability-specific estimates of score reliability) are easily estimated from the results obtained in IRT modeling (e.g., the 1-, 2-, or 3-parameter IRT models). Such information is useful for enhancing the precision of score reliability across different levels of ability or achievement (AERA, APA, NCME, 2014; Raju, Price, Oshima, & Nering, 2007). However, we acknowledge that CSEMs and conditional reliability estimates are less useful for or applicable to attitudinal or observational measurement scales often employed in nursing and/or healthcare research.

We applaud Sijsma and van der Ark's clear exposition of the factor analytic model and its utility for providing a framework for unifying reliability and validity. We really like factor analytic models and have published factor analytic-based techniques for estimating reliability (Gajewski et al., 2014). As with the CTT-based approach to reliability estimation, we do not want to be constrained only to the point estimates. We also need to ascertain the uncertainty in this point estimate—usually in the form of a 95% confidence interval—or as the Bayesians refer to as a “credible interval” (Jiang et al., 2014). The intervals provide us a level of uncertainty that reflects the sample size. For example, fitting a seven-item single factor (called *effort*) model (Gajewski et al., 2014, p. 96), we were able to estimate “entire reliability” using a bootstrap resample approach for estimating a 95% confidence interval (see Raykov & Shrout, 2002, p. 206), which was $R = .88$ (95% CI [.81, .92]). This is very similar to a calculation using the graphical user interface version of Bayesian instrument development (Jiang et al., 2014) that uses no prior information, and resulted in a calculation of $R = .87$ (95% CI [.84, .90]), very close to the bootstrap results. In both the bootstrap and the Bayesian approaches, we see that the intervals are both above .80, telling us that we are “95% sure” that the reliability of the items in the factor *effort* is above a common cutoff of 0.80.

Narrower intervals can be provided in the presence of an informative prior distribution that we argue can be provided by content validity data (Gajewski et al., 2012). Adding this prior information to the point and interval estimation process is the biggest advantage that the Bayesian approach offers over the bootstrap approach. This is especially useful since nurse researchers developing instruments recognize sample size as “especially critical when the researcher is examining disease entities that occur in small numbers in the population” (Pett, Lackley, & Sullivan, 2003, p. 48).

We programmed a beta version of user-friendly version of Bayesian instrument development (BID; a screen shot is shown in Figure 1) for R and WinBUGS. We demonstrated the new BID software for four clinical researchers. Existing clinical data from both content experts and participants were placed into the new software. Interactively with the graphical user interface (GUI), outcomes of the multiple runs of the data were reviewed by methodologists and clinical experts and were aligned with review group's expectations. The new BID software was readily usable by nurse researchers who were nonstatisticians. We are in the process of programming a next version and disseminating this software to nurse researchers.

Regarding a different point, Sijtsma and van der Ark's presentation of generalizability theory (Brennan, 1983; 2010) and its utility as a statistical model that provides powerful techniques for capturing multiple forms of error variance attributable to persons, items, and raters/observers is right on target. In our opinion, generalizability theory is underutilized in nursing and healthcare research (and many other fields) for: (a) reliability estimation using an existing instrument with multiple sources of error variance particular to raters or observers in addition to persons and items; and (b) instrument development and establishing evidence of psychometric adequacy. Finally, generalizability theory (Brennan, 1983; 2010) also provides a framework for linking or unifying reliability and validity in a single analysis. For example, in addition to reliability estimation, researchers can establish evidence for convergent and discriminant score validity through application of the multitrait-multimethod (MTMM) technique (Campbell & Fiske, 1959; Crocker & Algina, 1986; Price, 2015).

Despite the above, we enthusiastically endorse Sijtsma and van der Ark's clear exposition, and we applaud their promotion of model-based estimates of reliability as part of the process of validating instruments used in nursing research and health studies generally. Indeed, the authors have provided a cogent review of the three foundational models and methods of reliability estimation that have proven useful over time. The authors also point out several misconceptions that are commonly held by researchers regarding the estimate produced when using the classical true score model for estimating score reliability. For example, the authors note that the CTT equation is a tautology; then they review the assumptions of the CTT model of parallel tests, tau-equivalence, and the frequently misused "cutoff" point-estimate based criteria as evidence for an acceptable level of score or value reliability.

Acknowledgments

The authors acknowledge that research reported in this publication was supported by the National Institute of Nursing Research of the National Institutes of Health under Award Number R03NR013236. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). Standards for educational and psychological testing. American Educational Research Association; Washington, DC: 2014.
- Brennan, RL. Elements of generalizability theory. The American College Testing Program; Iowa City, IA: 1983.
- Brennan, RL. Generalizability theory. Springer; New York, NY: 2010.
- Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*. 1959; 56:81–105. doi:10.1037/h0046016. [PubMed: 13634291]
- Crocker, L.; Algina, J. Introduction to classical and modern test theory. Holt, Rinehart, and Winston; New York, NY: 1986.
- de Ayala, RJ. The theory and practice of item response theory. Guilford Publications; New York, NY: 2009.
- Gajewski BJ, Coffland V, Boyle DK, Bott MJ, Price LR, Leopold J, Dunton N. Assessing content validity through correlation and relevance tools: A Bayesian randomized equivalence experiment. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2012; 8:81–96. doi:10.1027/1614-2241/a000040.

- Gajewski BJ, Jiang Y, Yeh H-W, Engelman K, Teel C, Choi WS, Daley CM. Teaching confirmatory factor analysis to non-statisticians: A case study for estimating the composite reliability of psychometric instruments. *Case Studies in Business, Industry & Government Statistics*. 2014; 5:88–101. <http://revuessfds.math.cnrs.fr/index.php/csbig/article/view/218/202>.
- Jiang Y, Boyle DK, Bott MJ, Wick JA, Yu Q, Gajewski BJ. Expediting clinical and translational research via Bayesian instrument development. *Applied Psychological Measurement*. 2014; 38:296–310. doi:10.1177/0146621613517165. [PubMed: 24882893]
- Pett, MA.; Lackley, NR.; Sullivan, JJ. Making sense of factor analysis: The use of factor analysis for instrument development in health care research. Sage; Thousand Oaks, CA: 2003.
- Price, LR. Psychometric methods: Theory into practice. Guilford; New York, NY: 2015.
- Raju NS, Price LR, Oshima TC, Nering ML. Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*. 2007; 31:169–180. doi: 10.1177/0146621606291569.
- Raykov T, Shrout PE. Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*. 2002; 9:195–212. doi: 10.1207/S15328007SEM0902_3.
- Sijtsma K, Van der Ark LA. Conceptions of reliability revisited, and practical recommendations. *Nursing Research*. 2015; 64:xx–xx.

FIGURE 1.

Graphical user interface (GUI) using the Bayesian instrument development (BID) approach to fit a one-factor model. Several options are available in the GUI. “*Prior*” allows the use of prior information (True/False). “*Expertdata*” has a file of expert data responses for each of the items. The “*factor list*” indicates what subscale each item belongs to; in this case each of the seven items belongs to the first subscale (1). The “*EqualSpace*” is a condition of how the prior from the experts’ data is converted to loadings. The “*Participantdata*” is the usual spreadsheet of data from participants. Running the module returns a point estimate and 95% interval for reliability.