The many issues in reliability research: Choosing from a horn of plenty

Sijtsma, K.; van der Ark, L.A.

DOI
10.1097/NNR.0000000000000081
Publication date
2015
Document Version
Final published version
Published in
Nursing research

The Many Issues in Reliability Research

Choosing From a Horn of Plenty

Klaas Sijtsma ▼ L. Andries van der Ark

We respond to three commentaries on our discussion article on different conceptions of test score reliability. First, we discuss the use of standard errors for reliability estimates. Second, we discuss the desirability not to confuse issues pertaining to the dimensionality of the test data (closely related to construct validity) and the degree to which measurement values are repeatable under the same circumstances (i.e., the reliability issue). Third, we discuss a new reliability estimation method that is almost unbiased irrespective of the dimensionality of the test data.

Key Words: dimensionality of test data • divisive latent class reliability coefficient • psychometrics • reliability • standard error for reliability estimate • test score reliability

Nursing Research, March/April 2015, Vol 64, No 2, 152–154

We thank the discussants (Barbaranelli, Lee, Vellone, & Riegel, 2015; Gajewski, Price, & Bott, 2015; Yang & Green, 2015) for their excellent commentaries on our discussion article (Sijtsma & van der Ark, 2015). We discussed three reliability conceptions in an effort to guide readers of *Nursing Research* through the plethora of contributions to reliability theory. Together, these contributions offer a well-filled tool kit, but the wealth of different methods may also cause confusion about which tool to use. We argued that many researchers may be unaware that different reliability methods have conceptually different points of departure and that ignoring the different backgrounds may lead to an incorrect use of the methods. Remarkably, each discussion extended our exposition, and none showed overlap with the others, further illustrating the richness of the field. Although many topics in reliability research have been investigated, we believe that more topics need to be considered, and the discussants provide an anthology of additional, interesting issues. We briefly and, because of space limitations, incompletely reflect on the commentaries.

Precision of Reliability Estimates

Gajewski et al. (2015) addressed the important topic of precision of reliability estimates. Different random samples from the same population produce reliability estimates that differ because of sampling fluctuation, but many researchers tend to think of estimates based on one sample as population values and take a reliability estimate of, say, .82, as sufficient evidence that a threshold of .8 was exceeded (e.g., Yang & Green, 2015) and the test score was sufficiently reliable (Sijtsma, 2012). In general, researchers do not report standard errors or confidence intervals. We assume they do not use them and interpret estimates of reliability coefficients as if they were population values (Oosterwijk, van der Ark, & Sijtsma, 2012). Oosterwijk et al. found evidence that, for small samples and small numbers of items, uncertainty because of sampling fluctuation may be a more serious problem than the underestimation that most reliability methods produce; that is, under these circumstances, imprecision overrules bias. Hence, we applaud the efforts Gajewski et al. (2015) put into developing software addressing this problem in a Bayesian framework.

For coefficient alpha (Cronbach, 1951), several methods to compute standard errors and confidence intervals are available (e.g., Feldt, 1965; Kuijpers, van der Ark, & Croon, 2013; Maydeu-Olivares, Coffman, & Hartmann, 2007). By using a trick, the 95% confidence interval for coefficient alpha computed by Feldt's method can be obtained in software package SPSS (Statistical Package for the Social Sciences; IBM Corp., 2013). Coefficient alpha is a one-to-one function of the intraclass coefficient (e.g., Kistner & Muller, 2004), and consequently, the confidence interval for the intraclass coefficient may be used as a confidence interval for alpha. To obtain the 95% confidence interval, one needs to mark the checkbox "intraclass correlation coefficient" in SPSS's conventional reliability analysis. The confidence intervals proposed by Kuijpers et al. (2013) and Maydeu-Olivares et al. (2007) can be computed using R (R Core Team, 2014) and Mplus (Muthén & Muthén, 2010), respectively. We agree with Gajewski et al. (2015) that a conditional standard error of measurement available in an item response theory context offers better opportunities for addressing individuals' measurement precision than classical methods can provide (e.g., Mellenbergh, 1996).

Klaas Sijtsma, PhD, is Full Professor, Department of Methodology and Statistics, TSB, Tilburg University, The Netherlands.
L. Andries van der Ark, PhD, is Associate Professor, Research Institute for Child Development and Education, University of Amsterdam, The Netherlands.
DOI: 10.1097/NNR.0000000000000081

Dimensionality and Reliability

Barbaranelli et al. (2015) pointed out that knowledge of the scale's dimensionality is a necessary preliminary step in reliability analysis. We agree that, for the currently used reliability coefficients, knowledge of the scale's dimensionality is required to assess the degree of underestimation. For example, the more data deviate from unidimensionality, the larger the discrepancy between coefficient alpha and reliability. However, we emphasize that there are no compelling reasons why dimensionality and reliability would be related. Reliability answers the following question: "If I would repeat the measurement procedure under identical circumstances, to what degree would the measurement values be different from the values I have obtained in the first round?" In answering this question, dimensionality is not an issue. Of course, we advocate using substantively meaningful tests—that is, valid tests—but emphasize that one may be interested in the repeatability of the measurement irrespective of what the test measures, just as one may be interested in what the test measures irrespective of its degree of reliability. Failure to separate issues of meaning and interpretation (often related to dimensionality considerations) from repeatability expressed by formal, psychometric reliability is another source of confusion in discussions of psychometrics and the practical use of psychometric methods.

Yang and Green (2015) referred to this confusion when they noticed that hierarchical omega, ω_h, blurs the distinction between reliability and validity. This is what the factor analysis approaches to reliability do in general by separating a general factor from other factors and measurement error; the older thought model dividing measurement error into systematic error because of influences one did not wish to measure, and random measurement error is at the basis of the factor analysis approach to reliability. Of course, we do not propose to avoid the factor analysis approach to reliability as it indisputably provides illuminating insight into the composition of the test performance and how this may affect reliability, and researchers wishing to combine reliability and validity issues in one approach may have a point. Our stance is methodological, trying to separate dimensionality and repeatability issues from one another and encouraging researchers to take one step at a time so as to stay in control of two highly complex problems—reliability and validity.

For the sake of discussion about reliability and validity issues, we categorize both reliability and validity (defined as construct validity; e.g., Borsboom, van Heerden, & Mellenbergh, 2004) into "low" and "high" so that four combinations arise. The combination "high reliability, high validity" is desirable, and "low reliability, low validity" is undesirable. The combination "high reliability, low validity" refers to tests reliably measuring something one did not intend, for example, language skills (unwanted) in an arithmetic (wanted) test containing realistic context problems (e.g., arithmetic problems wrapped in a little story, like paying at the counter when shopping). Here, factor analysis may help to get a grip on the dimensionality of the data and improving the test by replacing items by better targeted items, possibly improving both reliability and validity. The combination "low reliability, high validity" may occur, for example, when complex, multidimensional job-performance assessment is right on target but is based on only one reviewer assessment. Here, a larger number of independent reviewers may help to increase reliability of the assessment.

On the basis of simulated data, we found that the recently proposed divisive latent class reliability coefficient (DLCRC; van der Palm, van der Ark, & Sijtsma, 2014; see also van der Ark, van der Palm, & Sijtsma, 2011) produces nearly unbiased estimates of reliability irrespective of the dimensionality in the data. Unlike other reliability coefficients for multidimensional data, the computation of DLCRC does not require that the dimensionality be specified beforehand. This result relativizes Barbaranelli et al.'s (2015) claim that knowledge of the scale's dimensionality is a necessary preliminary step in reliability analysis. We briefly outline method DLCRC.

First, the method estimates the joint density of the item scores by means of a divisive latent class model (van der Palm, 2013, Ch. 2) with K^* latent classes; for details about finding the optimal value for K^*, we refer the interested reader to van der Palm et al. (2014).

Let X_j denote the random variable for the score on item j ($j = 1,\ldots,J$), with realization x_j ($x_j = 0,\ldots, m$); then, $P(X_1 = x_1, \ldots, X_J = x_J)$ denotes the joint density of the J item scores. Let π_κ denote the weight of latent class κ, and let $P(X_j = x_j|\kappa)$ denote the probability that a respondent from latent class κ has score x_j on item j. In the (divisive) latent class model,

$$P(X_1{=}x_1, \ldots, X_J{=}x_J) = \sum_{\kappa=1}^{K^*} \pi_\kappa \prod_{j=1}^{J} P(X_j{=}x_j|\kappa). \quad (1)$$

Equation 1 breaks down the complex joint density of the item scores into smaller parts.

Second, using k as the second item index and y as the second item score, the DLCRC method uses the decomposition of the true-score variance as (Sijtsma & Molenaar, 1987)

$$\sigma_T^2 = \sum\sum_{j\neq k}\sum_x\sum_y P(X_j \geq x, X_k \geq y) - P(X_j \geq x)P(X_k \geq y)$$
$$+ \sum_j\sum_x\sum_y P(X_j \geq x, X_j \geq y) - P(X_j \geq x)P(X_j \geq y). \quad (2)$$

In Equation 2, the probability $P(X_j \geq x, X_j \geq y)$ refers to two replications of the same item j and thus is unobservable. All other probabilities are observable.

Third, σ_T^2 is estimated by replacing the observable probabilities in Equation 2 by their sample estimates and by replacing the unobservable probability $P(X_j \geq x, X_j \geq y)$ by the latent class estimator in Equation 1, that is, using g and h as item scores:

$$\tilde{P}(X_j \geq x, X_j \geq y)$$
$$= \sum_{g=x}^{m}\sum_{h=y}^{m}\sum_{\kappa=1}^{K^*} \pi_\kappa P(X_j{=}g|\kappa)P(X_j{=}h|\kappa). \quad (3)$$

A preliminary version of DLCRC is available from the R package mokken (van der Ark, 2012). We recommend using

DLCRC whenever a test is assumed to be multidimensional by design, that is, when the attribute to be measured is expected to be more complex than, for example, a simple skill or a well-defined, narrow ability. If the context arithmetic problems and the multidimensional job performance assessment represent the complex attribute the researcher wishes to measure using one test score, the researcher may consider using DLCRC. However, the examples clarify the reliability–validity issue well, and the first question the researcher needs to address is whether one test score based on a composite of subattributes is desirable or a score pattern based on approximately unidimensional subattributes that are measured separately. For the first option, DLCRC is suited, but for the second option, each subattribute requires a separate reliability estimate.

Accepted for publication December 9, 2014.

The authors declare no conflicts of interest.

Corresponding author: Klaas Sijtsma, PhD, Department of Methodology and Statistics, TSB, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands (e-mail: k.sijtsma@tilburguniversity.edu).

REFERENCES

Barbaranelli, C., Lee, C. S., Vellone, E., & Riegel, B. (2015). The problem with Cronbach's alpha…Comment on the Article by Sijtsma and Van der Ark. *Nursing Research*, *64*, 140–145.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty. *Psychometrika*, *30*, 357–370. doi:10.1007/BF02289499

Gajewski, B., Price, L. R., & Bott, M. (2015). Response to Sijtsma & van der Ark (2015): "Conceptions of reliability revisited, and practical recommendations. *Nursing Research*, *64*, 137–139.

IBM Corp. (2013). *IBM SPSS for Windows, version 22.0* [computer software]. Armonk, NY: Author.

Kistner, E. O., & Muller, K. E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika*, *69*, 459–474.

Kuijpers, R. E., van der Ark, L. A., & Croon, M. A. (2013). Testing hypotheses involving Cronbach's alpha using marginal models. *British Journal of Mathematical and Statistical Psychology*, *66*, 503–520.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*, 157–176. doi:10.1037/1082-989X.12.2.157

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299. doi:10.1037/1082-989X.1.3.293

Muthén, L. K., & Muthén, B. O. (2010). *Mplus version 6.1* [computer software]. Los Angeles, CA: Author.

Oosterwijk, P. R., van der Ark, L. A., & Sijtsma, K. (2012). *On the precision of reliability estimates*. Paper presented at the 77th International meeting of the Psychometric Society, Lincoln, NE.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, *77*, 4–20. doi:10.1007/s11336-011-9242-4

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, *52*, 79–97.

Sijtsma, K., & van der Ark, L. A. (2015). Conceptions of reliability revised, and practical recommendations. *Nursing Research*, *64*, 128–136.

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. Retrieved from http://www.jstatsoft.org/v48/i05/paper

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*, 380–392. doi:10.1177/0146621610392911

van der Palm, D. W. (2013). Latent class models for density estimation, with applications in missing data imputation and test-score reliability estimation [Unpublished doctoral dissertation]. Tilburg University, The Netherlands. Retrieved from http://www.dvdpalm.nl/thesis.pdf

van der Palm, D. W., Van der Ark, L. A., & Sijtsma, K. (2014). A flexible latent class approach to estimating test-score reliability. *Journal of Educational Measurement*, *51*, 339–357. doi:10.1111/jedm.12053

Yang, Y., & Green, S. B. (2015). Further discussion on reliability: The art of reliability estimation. *Nursing Research*, *64*, 146–151.