The Present and Future of Reliability Analysis

Advances in Theory and Practice

Julius Pfadt

Ulm University

October 26, 2021

Outline

Outline

Introduction

Reliability analysis:

- A quantification of measurement error
- How well does a test instrument capture systematic influences $\Leftrightarrow$ How repeatable is the measurement
- For multiple test administrations $\rightarrow$ measures of agreement (e.g., ICCs)
- For single test administrations $\rightarrow$ measures of consistency/repeatability (e.g., coefficient α) $\rightarrow$ current project

Classical Test Theory

- CTT defines classical reliability
- Split test score X_i of participant i into a hypothetical true part T_i and an error part E_i
- On a test score level:

$$X = T + E \tag{1}$$

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{2}$$

Reliability

Reliability ρ equals the correlation of parallel tests:

$$\rho = \rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \tag{3}$$

- True scores T and T' of parallel tests correlate 1 per definition
- Error scores E and E' of parallel tests correlate 0 per definition

$\rightarrow$ Reliability answers the question how likely it would be to see the same results if the test was readministered

CTT-Coefficients

- Decompose the data covariance matrix $\mathbf{\Sigma}$ trying to disentangle true score variance from error score variance
- Popular coefficients:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\mathrm{tr}(\mathbf{\Sigma})}{\mathbf{\Sigma}}\right) \tag{4}$$

$$\lambda_2 = \frac{\mathbf{\Sigma} - \mathrm{tr}(\mathbf{\Sigma}) + \sqrt{\frac{k}{k-1}\,c}}{\mathbf{\Sigma}} \tag{5}$$

$$\lambda_4 = \max\left[2 * (1 - \frac{\sigma_A^2 + \sigma_B^2}{\sigma_X^2})\right] \tag{6}$$

$$\mathrm{glb} = 1 - \frac{\mathrm{tr}(\mathbf{\Sigma}_E)}{\mathbf{\Sigma}} \tag{7}$$

Factor Analysis

- Split test score X_i of participant i into a part explained by one or more factors F_i and a part that cannot be explained, E_i:

$$X = \Lambda F + E \tag{8}$$

- Reliability is the relative amount of test score variance that can be explained by the factor(s):

$$\rho = \frac{\sum \Lambda^2}{\sigma_X^2} \tag{9}$$

- True score variance is replaced by the factor explained variance
- The adequacy of the reliability approximation is now dependent on the fit of the factor model

FA-Coefficients

- Unidimensional data → based on single-factor model:

$$\omega_u = \frac{(\sum \lambda)^2}{(\sum \lambda)^2 + \sum \psi} \tag{10}$$

- Multidimensional data → based on bi-factor model:

$$\omega_t = \frac{\sum \Lambda^2}{\sum \Lambda^2 + \sum \psi} \tag{11}$$

$$\omega_h = \frac{(\sum \lambda_g)^2}{(\sum \lambda_g)^2 + \sum \psi}\,. \tag{12}$$

- ω_t estimates total reliability, ω_h estimates g-factor reliability

Outline

Coefficient α (and other CTT-Coefficients)

Properties:

- Coefficient α equals the reliability when test items satisfy true-score equivalence (e.g., Lord & Novick, 1968)
- Coefficient α is smaller than the reliability when true-score equivalence is violated $\rightarrow$ lower bound (e.g., Sijtsma, 2009)
- Discrepancy increases with multidimensionality (e.g., Dunn et al., 2014)

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*. https://doi.org/10.1007/s11336-021-09789-8

Coefficient α Discussion

Criticism (1): "Essential true-score equivalence is unrealistic; hence, lower bounds must not be used"

Counter-argument (1): "All models are wrong"

- Models are perfect descriptions of an imperfect reality $\rightarrow$ fit by approximation
- We accept a certain amount of misfit for FA coefficients
- When true-score equivalence does not hold $\rightarrow$ coefficient α becomes a lower bound

Coefficient α Discussion

Counter-argument (2): Lower bounds are useful in practice

- Conservative estimation is desired in high stake conditions (admissions test, medical diagnosis)
- With unidimensional data, the discrepancy of lower bounds is generally negligible (see, e.g., Hunt & Bentler, 2015)
- With multidimensional data, unidimensional subsets can be used
- Contrary to FA, CTT is tautological, meaning $X = T + E$ is always true, and the CTT-coefficients are always lower bounds to the reliability

Coefficient α Discussion

Criticism (2): "Correlated errors cause the failure of the lower bound theorem" $\rightarrow$ *Coefficient* α *may be larger than the reliability*

Counter-argument (1): CTT and FA approaches are conceptually different

- Correlated errors are associated with non-target influences
- FA methods try to disentangle target from non-target influences and define reliability based on the target-influences only
- CTT methods try to indicate the degree to which a measurement is repeatable under the same circumstances $\rightarrow$ non-target influences that are repeatable are included in the true score

$\rightarrow$ CTT and FA define different forms of reliability

Coefficient α Discussion

Counter-argument (2): "The lower bound theorem assumes uncorrelated errors"

- The use of CTT assumes that errors are uncorrelated, because everything systematic is part of the true score
- Assuming correlated errors means leaving CTT and reliability defined by CTT $\rightarrow$ the lower bound theorem is invalid
- The same test can have multiple reliabilities, since in CTT reliability is always dependent on test-group-procedure

Pfadt, J. M., & Sijtsma, K. (2021). *Statistical properties of lower bounds and factor analysis methods for reliability estimation*. [Manuscript submitted for publication]

Simulation Study

Background:

- Previous simulation studies used a factor model to generate the data
- The reliability equals the FA reliability coefficient
- Coefficient α is compared to FA coefficients
- Does the data generation affect the performance of the reliability coefficients?

$\rightarrow$ New simulation study with two types of data generation models

Conditions:

- Data generation: From an IRT and a FA-model; unidimensional and multidimensional
- $k = 9/18$, $n = 500/2000$
- Separate condition with a misspecified model
- Coefficients: α, λ_2, λ_4, glb, ω_u, ω_h, ω_t

Results: Unidimensional Data

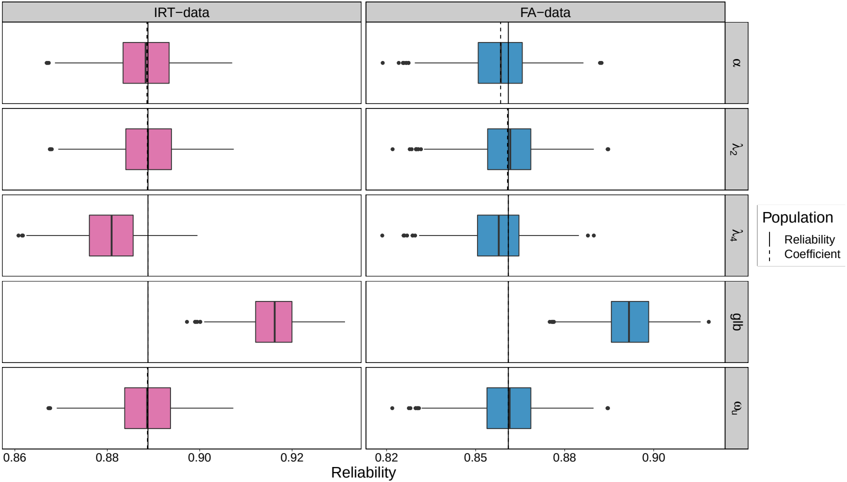


Figure 1. The point estimates of the coefficients across 1,000 simulation runs for $k = 18$ items and sample size of $n = 500$. In the IRT-conditions the data were generated from a 2-parameter graded response model. In the FA-conditions the data were generated from a single-factor model.

Results: Multidimensional Data

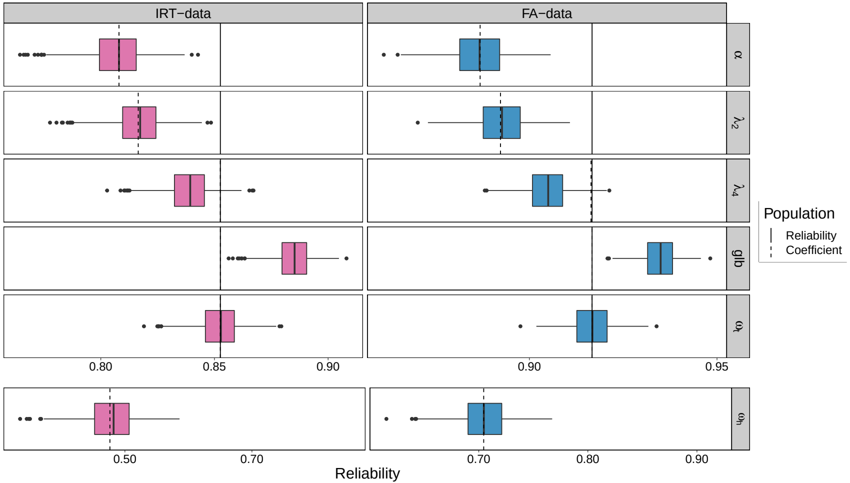


Figure 2. The point estimates of the coefficients across 1,000 simulation runs for $k = 18$ items and sample size of $n = 500$. In the IRT-conditions the data were generated from a 2-parameter graded response model with three latent variables and intercorrelations of .3. In the FA-conditions the data were generated from a second-order factor model with three primary latent variables.

Misspecified Models

Case (1):

- Population model is multidimensional with a common factor
- Analysis assumed unidimensionality $\rightarrow$ estimated coefficient ω_u

Case (2):

- Population model is purely multidimensional with no common factor
- Analysis assumed a common factor $\rightarrow$ estimated coefficients α, λ_2, λ_4, glb, ω_h, ω_t

Results: Misspecified Models

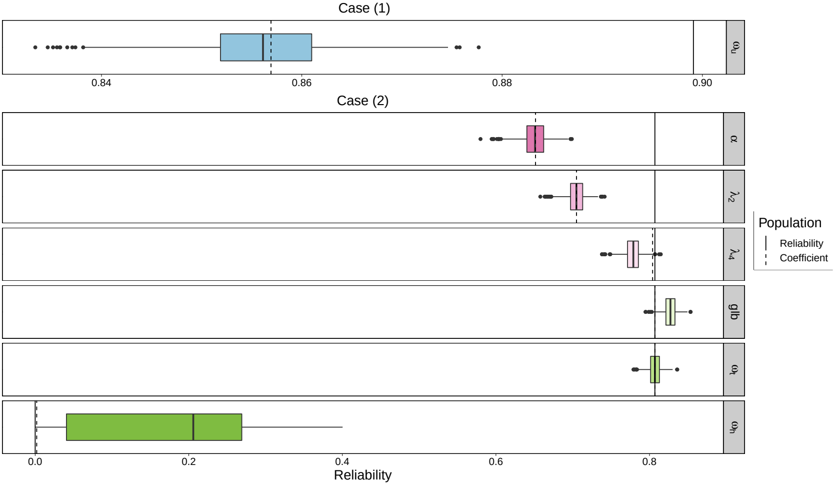


Figure 3. The point estimates of the coefficients across 1,000 simulation runs with $n = 1,000$. The data for Case (1) was generated from a second-order factor model with three primary latent variables. The data for Case (2) was generated from a factor model with three latent variables and no intercorrelations.

Simulation Study

Results summary:

- No meaningful differences between the IRT and FA conditions
- With unidimensional data, most coefficients performed well
- With multidimensional data, the lower bounds performed unsatisfactory
- Coefficient λ_2 was at least as good as α
- The ω-coefficients performed well

Conclusions:

- When data are unidimensional $\rightarrow$ use any reliability coefficient except the glb
- When data are multidimensional and the total reliability is of interest $\rightarrow$ use ω_t
- When using an FA-coefficient $\rightarrow$ confirm model fit

Outline

Uncertainty Estimation

- Account for sampling error by indicating the uncertainty of a parameter point estimate with, e.g., a standard error or an interval
- In reliability reporting, this practice is virtually non existent (Flake et al., 2017; Moshagen et al., 2019; Oosterwijk et al., 2019)
- Possible reasons for this:
  - Reliability is a “minor” analysis
  - Intervals are contrary to the idea of reliability cutoffs
  - The idea that reliability as an indication of measurement error is prone to sampling error is overlooked

“There is no excuse whatever for omitting to give a properly determined standard error [...]. All statisticians will agree with me here, [...].” (Jeffreys, 1961, p. 410)

Confidence Intervals

- The 95% confidence interval covers the parameter in 95% of the cases when one would repeat the process of sampling and computing the 95% confidence interval for the parameter numerous times (Morey et al., 2016; Neyman, 1937).

- Misconception: "The 95% confidence interval of a parameter contains the parameter with 95% probability; one can be 95% certain that the interval contains the parameter."

→ A 95% **credible interval** contains the parameter with 95% probability

Bayesian Parameter Estimation

$$\Pr(\theta|D) \propto \Pr(D|\theta)\Pr(\theta)\,. \tag{13}$$

- Combine the prior distribution of a parameter, $\Pr(\theta)$, with the likelihood of the data given the parameter, $\Pr(D|\theta)$, to yield the posterior distribution of the parameter, $\Pr(\theta|D)$.
- The prior distribution contains the probabilities for all parameter values before observing the data D
- The posterior distribution of the parameter contains the probabilities of all parameter values after observing the data

Bayesian Parameter Estimation: Visualization

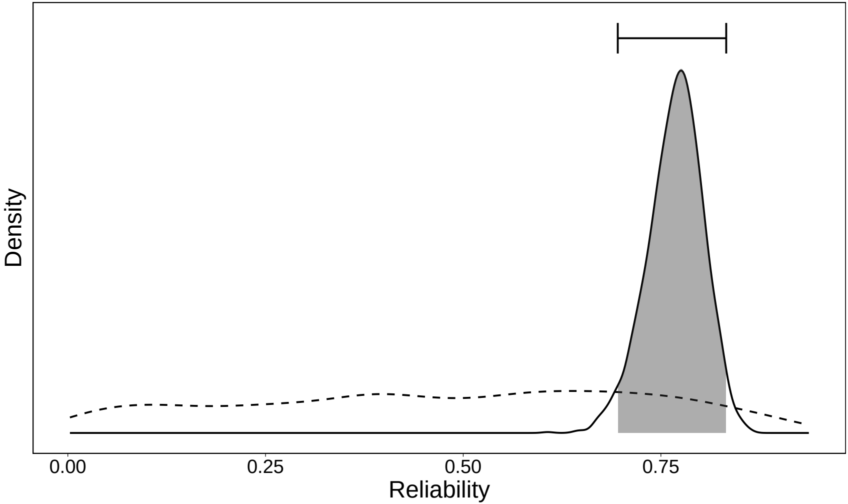


Figure 4. An example prior-posterior plot of coefficient α. The dotted line denotes the prior distribution, the straight line the posterior distribution. The error bar and the gray area denote the 95% credible interval.

Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E.-J. (2021). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*, 1–30. https://doi.org/10.1080/00273171.2021.1891855

Goal

Obtain the posterior distributions of the reliability coefficients for unidimensional data (coefficients α, λ_2, glb, and ω_u):

- $\rightarrow$ Obtain Bayesian point estimates and credible intervals
- $\rightarrow$ Answer questions such as: "How likely is it that the reliability of this test is higher than .80?"
- $\rightarrow$ Incorporate prior knowledge into the analysis

Bayesian Estimation

We distinguish two groups of coefficients:

CTT-coefficients:

- α, λ_2, glb
- Calculated from the data covariance matrix

→ Estimate the covariance matrix in the Bayesian framework

→ Compute the posterior distributions of the reliability coefficients from the posterior distribution of the covariance matrix

FA-coefficient:

- ω_u
- Estimated from the data matrix by fitting a single-factor model

→ Estimate the single-factor model in the Bayesian framework

→ Compute the posterior distribution of ω_u from the posterior distributions of the single-factor model parameters

Bayesian Estimation

CTT-Coefficients:

- Both the prior and posterior distribution of the covariance matrix are an inverse Wishart distribution when the data follow a multivariate normal distribution (Murphy, 2007)
- We sample numerous times (e.g., 2,000) from the inverse Wishart with hyperparameters based on the data
- We obtain a posterior sample of covariance matrices that are an adequate representation of the posterior distribution
- We compute posterior samples of the CTT-coefficients using equations (4), (5), (7) from the posterior sample of covariance matrices

Bayesian Estimation

FA-coefficient:

- Borrow the prior distributions for the single-factor model parameters from Bayesian structural equation modeling (see Lee, 2007):
  - A normal distribution for the factor loadings and the factor scores
  - An inverse gamma distribution for the residuals
  - An inverse Wishart distribution for the covariance matrix of the latent variables
- We implement Markov chain Monte Carlo (MCMC) sampling to obtain posterior samples of loadings and residuals
- We compute the posterior samples of ω_u from the posterior samples of loadings and residuals

Simulation Study

How do the Bayesian reliability coefficients perform statistically?
$\rightarrow$ Compare them to classical point estimates and bootstrapped confidence intervals in a simulation study with multiple conditions:

- Data were generated from a single-factor model
- Number of items: 5 and 20
- Sample size: 50, 100, and 500
- Average inter-item correlation: 0, .3, and .7

The results included:

- Root mean square error of point estimates
- Coverage of 95% uncertainty intervals
- Probability of overestimation

Simulation Study

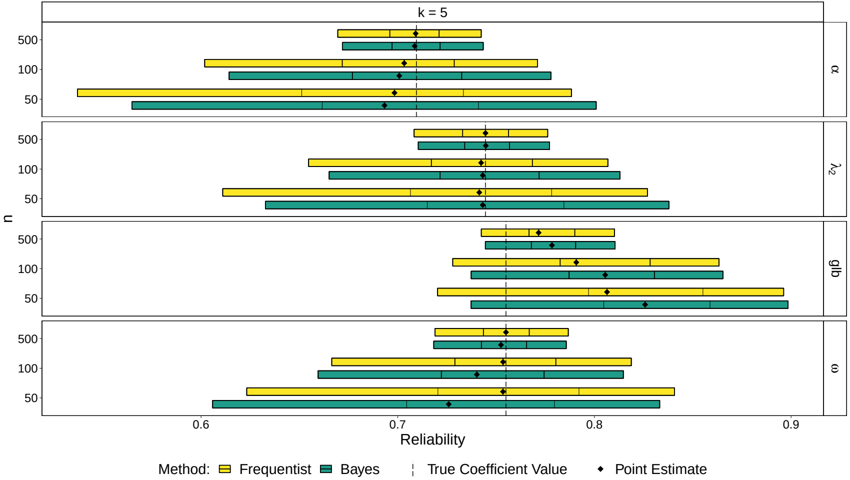


Figure 5. Simulation results for the medium-correlation condition with $k = 5$ items. The endpoints of the bars are the mean 95% uncertainty interval limits. The 25%- and 75%-quartiles are indicated with vertical line segments.

Simulation Study

Results summary:

- The credible intervals for coefficients α, λ_2, and ω_u performed satisfactory,
- The Bayesian point estimation was slightly worse than the classical (frequentist) in small samples
- The results for the classical bootstrap confidence intervals and the Bayesian credible intervals generally agreed

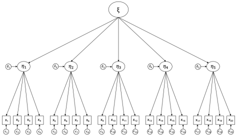
Conclusions:

- Use uncertainty estimates to accompany point estimates of α, λ_2, and ω_u, preferably the credible intervals we implemented
- The use of intervals is even more important when the sample size is small

Pfadt, J. M., van den Bergh, D., & Moshagen, M. (2021). *The reliability of multidimensional scales: A comparison of confidence intervals and a Bayesian alternative* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/d3gfs

Introduction

- Coefficients ω_t for the total reliability and ω_h for the g-factor reliability (see Equations 11 and 12)
- The ω-coefficients can be based on a second-order factor model:



  - relates several primary group factors to the items (facets, dimensions)
  - relates a general secondary factor to the group factors (common attribute)
  - is nested in the bi-factor model
- The second-order factor model loadings are transformed to yield the bi-factor model loadings for ω_t and ω_h

Motivation

- Credible intervals for coefficients ω_t and ω_h are not available
- Different methods to obtain confidence intervals of ω_t and ω_h are scarcely researched

$\rightarrow$ Develop Bayesian versions of ω_t and ω_h

$\rightarrow$ Compare multiple confidence intervals

Bayesian Estimation

- Similar to coefficient ω_u and the single-factor model
- Prior distributions for the second-order factor model (see Lee, 2007):
    - A multivariate normal distribution for the group factor loadings, and the factor scores
    - A normal distribution for the general factor loadings
    - An inverse gamma distribution for the manifest and the latent residuals
    - An inverse Wishart distribution for the covariance matrix of the latent variables
- We use MCMC sampling
- We compute the posterior samples of ω_t and ω_h from the posterior samples of loadings and residuals

Simulation Study

How do the Bayesian versions of ω_t and ω_h perform statistically? How do different confidence intervals perform?

Confidence intervals:

- EFA based non-parametric bootstrap intervals: Standard error (SE), standard error bias corrected (SE_{Bias}), standard error log transformed (SE_{Log}), percentile (Perc), bias corrected and accelerated (BCA)
- CFA based Wald-type interval (Wald)

Conditions:

- Data were generated from a second-order factor model
- Level of reliability: Low (.5) and high (.8)
- Number of items (model size): 9 (three group factors) and 30 (five group factors)

Results included:

- Root mean square error of point estimates
- Coverage of 95% uncertainty intervals

Simulation Study: Coverage Results

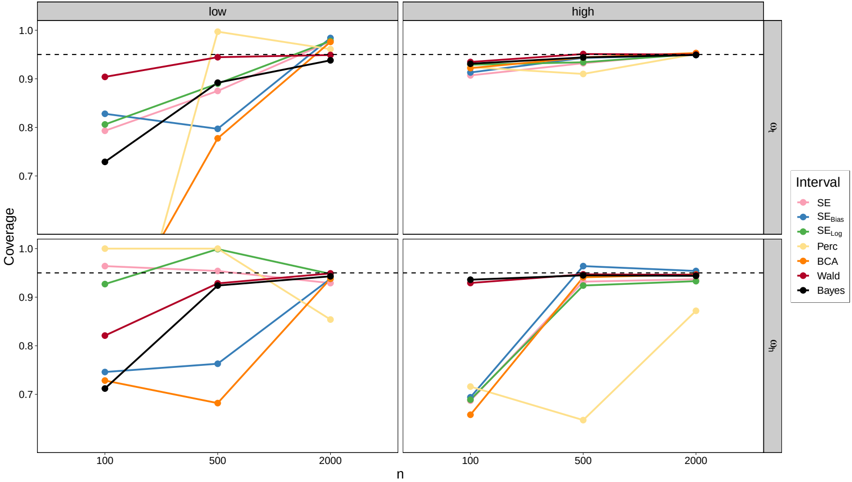


Figure 7. The coverage of the confidence intervals (SE - Wald), and the credible interval (Bayes) for $k = 30$ items and five group factors. The closer the dots are at the 95% line the better. Low reliability equaled $\omega_t = .5$; high reliability equaled $\omega_t = .8$.

Simulation Study

Results summary:

- Out of the confidence intervals, the SE, SE_{Log}, and Wald interval performed best
- The credible intervals performed satisfactory in most conditions
- With small samples and low reliability none of the intervals performed well

Conclusions:

- Use intervals for ω_t and ω_h, preferably credible intervals
- Be cautious with multidimensional reliability estimation when sample size is small and the reliability low
- Out of the confidence intervals, we recommend the Wald-type interval if the CFA converges

Bridging the Gap between Theory and Practice

- We implemented all methods in the R-package `Bayesrel`
- The R framework addresses researchers familiar with programming
- For others, the use of the Bayesian reliability estimates depends on an implementation in GUI-based software, such as SPSS

$\rightarrow$ **JASP**:

- Statistical click-and-response program much like SPSS but free of charge
- Developed by a team around EJ Wagenmakers at the University of Amsterdam
- Offers many popular analyses in a classical and a Bayesian way

Pfadt, J. M., van den Bergh, D., Sijtsma, K., & Wagenmakers, E.-J. (2021). *A tutorial on Bayesian single-test reliability analysis with JASP* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/j6z8h

Tutorial

- Data set from Nicolai and Moshagen (2018) containing a self-rating scale for manic symptoms (ASRM) from 78 participants
- Complete Bayesian reliability analysis in JASP with coefficients ω_u and α:
  - Point estimates and credible intervals
  - Prior-posterior plots
  - Probability that coefficient is higher than, e.g., .70
  - If-item-dropped statistics and plots
  - Assessing convergence
  - Checking model fit with the posterior predictive check
  - Missing data handling

$\rightarrow$ small demo...

Outline

Limitations

- CTT is a quite rudimentary measurement model, CTT-reliability might not be what people wish to know (target vs. non-target influences)
- Simulation studies are based on ideal situations (multivariate normal data, perfectly fitting models), real data is messy
- Relatively uninformative priors were used, other priors might yield different results
- Priors were never set on the reliability parameters itself but the covariance matrix and the factor model parameters

Conclusions

Psychometric models:

- Coefficient α is a lower bound to the reliability as defined by CTT
- Psychological theory $\rightarrow$ measurement model $\rightarrow$ type of reliability $\rightarrow$ reliability coefficient
- With unidimensional data, the choice of a reliability coefficient is almost arbitrary
- With multidimensional data, the FA-coefficients are recommended

Uncertainty estimation:

- Use intervals for reliability estimates!
- Credible intervals for reliability coefficients are highly practical and – through this work – accessible
- Some confidence intervals for reliability coefficients perform well and should accompany classical reliability point estimates

Contributions

- Discussion of two recurring criticisms of coefficient α, showing α is a useful lower bound to reliability defined by CTT
- Comparison of multiple popular CTT and FA reliability coefficients using different data generating models
- Investigation of different confidence intervals for popular reliability coefficients
- Development of popular CTT and FA reliability estimates for unidimensional and multidimensional data in the Bayesian framework
- Implementation of all developed methods – Bayesian and classical – in the R-package `Bayesrel` and in JASP for a wide audience to use

Thank you for your attention!

References I

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Hunt, T. D., & Bentler, P. M. (2015). Quantile lower bounds to reliability based on locally optimal splits. *Psychometrika*, *80*(1), 182–195. https://doi.org/10.1007/s11336-013-9393-6

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. John Wiley & Sons Ltd. https://doi.org/10.1002/9780470024737

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Moshagen, M., Thielmann, I., Hilbig, B. E., & Zettler, I. (2019). Meta-analytic investigations of the HEXACO Personality Inventory(-Revised): Reliability generalization, self-observer agreement, intercorrelations, and relations to demographic variables. *Zeitschrift für Psychologie*. https://doi.org/10.1027/2151-2604/a000377

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution*. University of British Columbia.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A*, *236*(767), 333–380. https://doi.org/10.1098/rsta.1937.0005

Nicolai, J., & Moshagen, M. (2018). Pathological buying symptoms are associated with distortions in judging elapsed time. *Journal of Behavioral Addictions*, *7*(3), 752–759. https://doi.org/10.1556/2006.7.2018.80

Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2019). Using confidence intervals for assessing reliability of real tests. *Assessment*, *26*(7), 1207–1216. https://doi.org/10.1177/1073191117737375

Pfadt, J. M., & Sijtsma, K. (2021). *Statistical properties of lower bounds and factor analysis methods for reliability estimation*. [Manuscript submitted for publication].

References II

Pfadt, J. M., van den Bergh, D., & Moshagen, M. (2021). *The reliability of multidimensional scales: A comparison of confidence intervals and a Bayesian alternative* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/d3gfs

Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E.-J. (2021). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*, 1–30. https://doi.org/10.1080/00273171.2021.1891855

Pfadt, J. M., van den Bergh, D., Sijtsma, K., & Wagenmakers, E.-J. (2021). *A tutorial on Bayesian single-test reliability analysis with JASP* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/j6z8h

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*. https://doi.org/10.1007/s11336-021-09789-8