The Present and Future of Reliability Analysis

Advances in Theory and Practice

Julius Pfadt

Thesis Defense at Ulm University

March 14, 2023

Outline

1. Reliability

2. Part I: The Choice of Coefficients
   - Article I: Two Recurring Criticisms of Coefficient α: A Discussion of Lower Bounds and Correlated Errors
   - Article II: Coefficient α and the Future of Reliability: A Rejoinder
   - Article II: Statistical Properties of Lower Bounds and Factor Analysis Methods for Reliability Estimation

3. Part II: The Choice of Estimation
   - Article IV: Bayesian Estimation of Single-Test Reliability Coefficients
   - Article V: A Tutorial on Bayesian Single-Test Reliability Analysis with JASP
   - Article VI: Classical and Bayesian Uncertainty Intervals for the Reliability of Multidimensional Scales

4. Conclusions

Motivation

Measurement in psychology is not perfect

↓

Researchers try to quantify measurement error = reliability analysis

↓

How can the status quo be advanced?

↓

(1) Improve the understanding of popular reliability coefficients
(2) Improve the way these coefficients are estimated with new methods

Measurement in Classical Test Theory (CTT)

- 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 ρ:

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

Reliability in CTT

- A measurement instrument that is *reliable* yields similar results if administered to the same people multiple times
- For instance, a bathroom scale, or an intelligence test

- Classical definition of reliability: The repeatability of a measurement

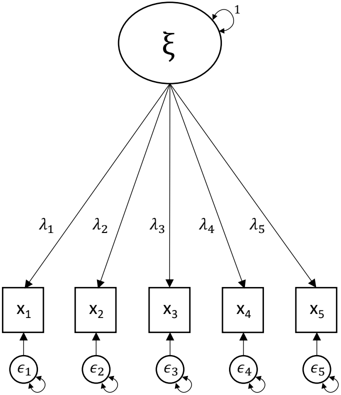
CTT-Reliability

Reliability ρ *equals the correlation of parallel tests:*

$$\rho = \rho_{XX'} \tag{4}$$

- Parallel tests X and X' are identical tests that are administered to the same sample of participants under the same conditions
- The correlation of parallel test scores equals the proportion of test score variance that is true score variance
- However, parallel tests are unavailable in practice
- CTT-coefficients approximate the reliability from a single test administration: α, λ_2, greatest lower bound (glb)

Another measurement theory: Factor analysis (FA)

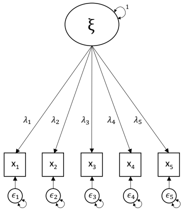


Factor Analysis

- Split test score X_i of participant i into a part explained by one or more factors F_i (latent variables) and a part that cannot be explained, E_i. Test score level:

$$X = \Lambda F + E \tag{5}$$

- Loadings Λ indicate how much influence the factor has on the item responses



FA-Reliability

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

$$\rho = \frac{\sum \Lambda^2}{\sigma_X^2} \tag{6}$$

- Reliability depends on the fit of the factor model
- FA-coefficients: ω_u for unidimensional data, ω_t and ω_h for multidimensional data

Outline

What coefficients should researchers choose to estimate reliability?

Coefficient α (and other CTT-Coefficients)

- Coefficient α equals reliability when test items are essentially true score equivalent (e.g., Lord & Novick, 1968)
- Coefficient α is smaller than the reliability when test items are not ess. true score equivalent $\rightarrow$ lower bound (e.g., Sijtsma, 2009)
- The more multidimensional a test the smaller coefficient α compared to the reliability (e.g., Dunn et al., 2014)

The use of coefficient α has been criticized a lot (Cho, 2016; Cho & Kim, 2015; Dunn et al., 2014; Graham, 2006; Green & Hershberger, 2000; Green & Yang, 2009; Lucke, 2005; Teo & Fan, 2013).

Article I

Sijtsma, K., & Pfadt, J. M. (2021a). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*(4), 843–860. https://doi.org/10.1007/s11336-021-09789-8

Coefficient α Discussion

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

Coefficient α Discussion

Counter-argument (1): All models are wrong

- Models are perfect descriptions of an imperfect reality $\rightarrow$ fit by approximation
- When true-score equivalence does not hold $\rightarrow$ coefficient α becomes a lower bound

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

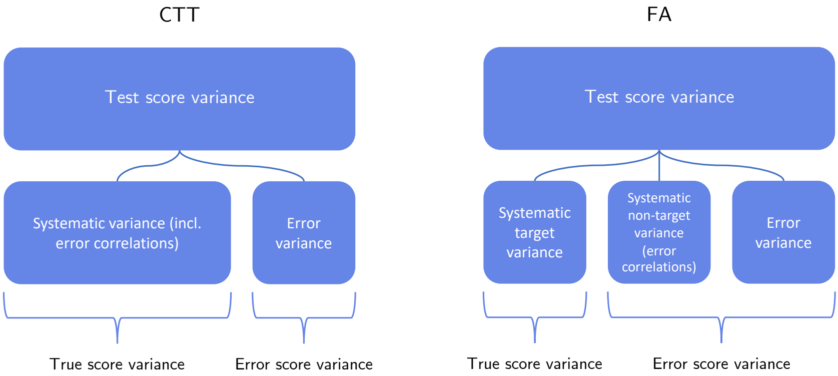
- Conservative estimation is desired in high stake conditions (admissions test, medical diagnosis)
- With unidimensional data, the discrepancy of lower bounds is generally small (see, e.g., Hunt & Bentler, 2015)
- CTT model always fits

Coefficient α Discussion

Criticism (2): "With correlated errors the lower bound property of coefficient α fails" $\rightarrow$ Coefficient α may be larger than the reliability

Coefficient α Discussion

Counter-argument: CTT and FA approaches are conceptually different



$\rightarrow$ CTT and FA define different reliabilities because they define the true score variance differently

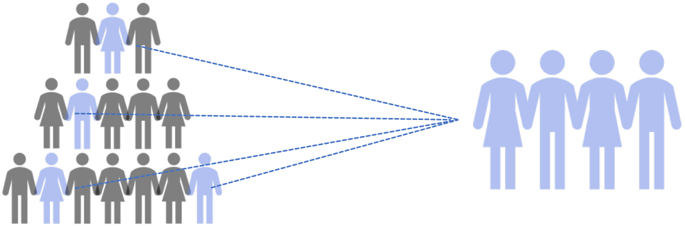
Coefficient α – Discussion

CTT and FA approaches are conceptually different

- CTT assumption: Errors are uncorrelated, because all systematic (repeatable) influences are part of the true score
- Assuming correlated errors means leaving CTT $\rightarrow$ properties derived from it become invalid (lower bound theorem)
- In CTT, reliability depends on test-group-procedure
- In FA, separating systematic non-target variance (correlated errors) tries to free reliability from the influence of the procedure

Outline

In practice, researchers report a coefficient α point estimate for their reliability analysis.



Uncertainty Estimation

> *"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)

- In psychological studies we draw a finite sample from a population $\rightarrow$ sampling error
- How to generalize the results to the population?
- Proper statistical practice: Account for sampling error by indicating the uncertainty of a parameter point estimate with, e.g., a standard error or an interval
- However, in reliability, this practice is virtually non existent (Flake et al., 2017; Moshagen et al., 2019; Oosterwijk et al., 2019)

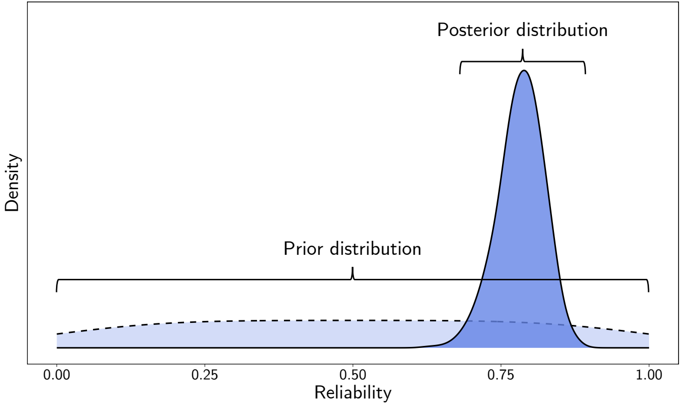
Frequentist Framework: Confidence Intervals

- 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."
- Probability if a specific reliability confidence interval covers the true parameter is unknown
- Definition: *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).

→ A 95% *credible interval* (Bayesian framework) contains the parameter with 95% probability

Bayesian Parameter Estimation

$$\overbrace{p(\theta \mid X)}^{\text{posterior}} \propto \overbrace{p(X \mid \theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}} \tag{7}$$



Bayesian Reliability Estimation

Benefits:

- Probability that the reliability parameter lies in a specific interval, for instance, the 95% credible interval
- Probability that the reliability exceeds a specific value, for instance, .80
- Incorporate prior knowledge about the reliability of a test instrument into the analysis

Obstacle: The posterior distributions of reliability coefficients are generally unavailable to researchers

How to obtain the posterior distributions of CTT and FA reliability coefficients?

Article IV

Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E.-J. (2022). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*, *57*(4), 620–641. https://doi.org/10.1080/00273171.2021.1891855

Article VI

Pfadt, J. M., van den Bergh, D., & Moshagen, M. (in press). Classical and Bayesian uncertainty intervals for the reliability of multidimensional scales. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2022.2124162

CTT-Coefficients (α, λ_2, glb)

- Calculated from the data covariance matrix

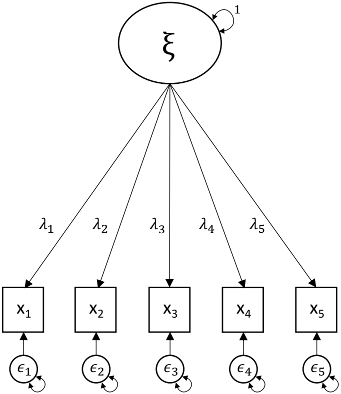
→ Estimate the covariance matrix in the Bayesian framework:
  - Data are multivariate normal
  - Conjugate prior for the covariance matrix: inverse Wishart distribution

  → sample directly from the posterior distribution of the covariance matrix, with hyperparameters obtained from the data (Gelman et al., 2013)

- From the posterior covariance matrices compute posterior samples of the CTT-coefficients using the coefficient formulas

FA-Coefficients – Unidimensional

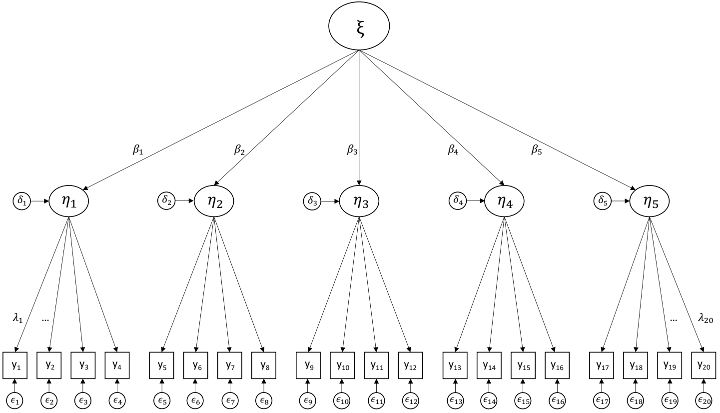
Coefficient ω_u:



Single-factor model

FA-Coefficients – Multidimensional

Coefficients ω_t and ω_h:



Second-order factor model

Bayesian Factor Model Estimation

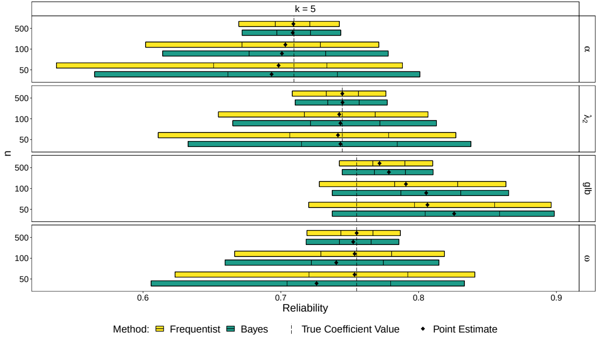
- Methodology from Bayesian SEM (Lee, 2007):
  - Data are multivariate normal
  - Conjugate priors: Normal distributions for loadings and factor scores, inverse gamma distributions for residual variances
- Posteriors via Gibbs sampling: Draw from the posterior distribution of a model parameter conditional on the remaining model parameters
- Using the posterior samples of loadings and residual variances compute the posterior samples of $\omega_u/\omega_t/\omega_h$ using the coefficient formulas

Simulation Studies

How do the Bayesian reliability coefficients perform statistically compared to confidence intervals? → Simulations with multiple conditions

Unidimensional results:

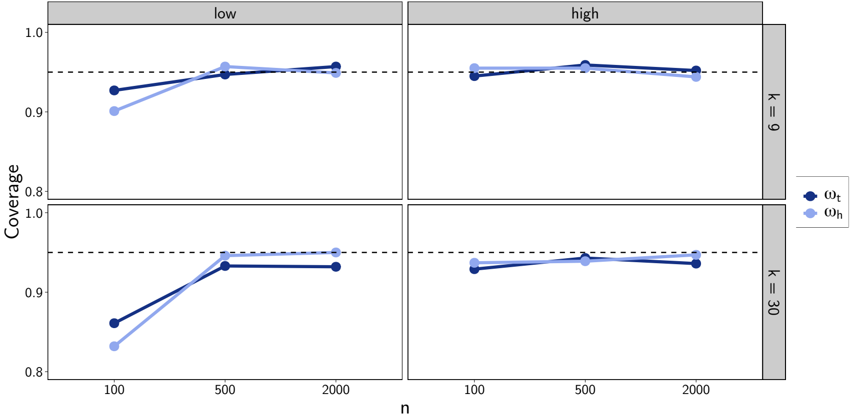
- Similar credible and confidence intervals
- The Bayesian versions of α, λ_2, glb, ω_u performed well across realistic conditions: Point estimates converged on the population values and coverage reached to .95



Simulation Studies

Multidimensional results:

- The Bayesian ω_t, ω_h performed well; however, with low reliability a relatively large sample size (N=500) was needed for satisfactory coverage



Simulation Studies – Conclusion

The Bayesian coefficients perform well and should be applied for uncertainty estimation in reliability.

Bridging the Gap between Theory and Practice: R

- The R-package `Bayesrel` contains all developed methods
- 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

Bridging the Gap: **JASP**

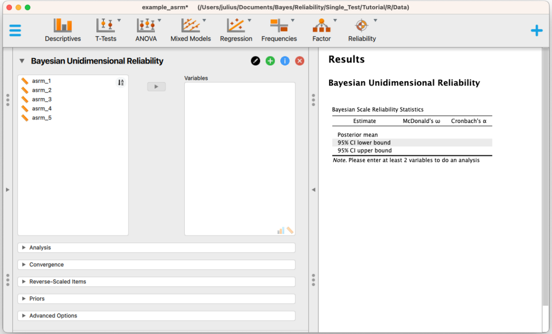
- Statistical click-and-response program much like SPSS but free and open-source
- Offers many popular analyses in a classical and a Bayesian way
- Perfect environment to implement Bayesian reliability estimates

Article V:

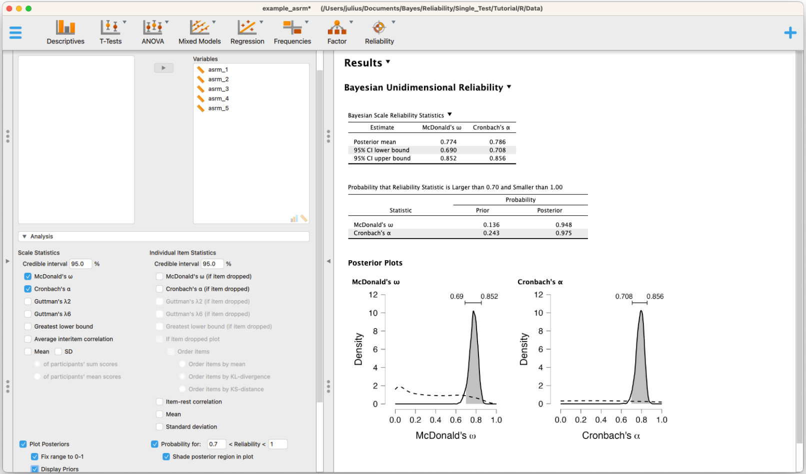
Pfadt, J. M., van den Bergh, D., Sijtsma, K., & Wagenmakers, E.-J. (in press). A tutorial on Bayesian single-test reliability analysis with JASP. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01778-0

Tutorial

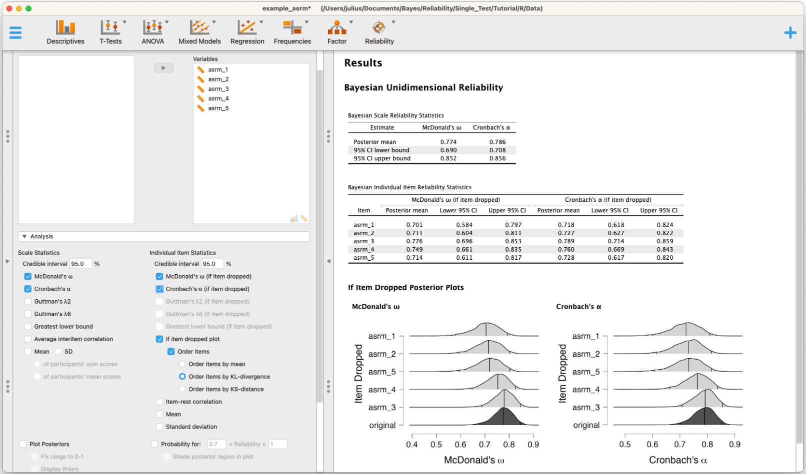
- Complete Bayesian reliability analysis in JASP with coefficients ω_u and α
- Data set from Nicolai and Moshagen (2018) containing the responses of 78 participants on a 5-item self-rating scale for manic symptoms (ASRM)



Tutorial


example_asrm* (/Users/julius/Documents/Bayes/Reliability/Single_Test/Tutorial/R/Data)
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Reliability
Variables
asrm_1
asrm_2
asrm_3
asrm_4
asrm_5
Analysis
Scale Statistics
Credible interval 95.0 %
McDonald's ω
Cronbach's α
Guttman's λ2
Guttman's λ6
Greatest lower bound
Average interitem correlation
Mean
SD
of participants' sum scores
of participants' mean scores
Individual Item Statistics
Credible interval 95.0 %
McDonald's ω (if item dropped)
Cronbach's α (if item dropped)
Guttman's λ2 (if item dropped)
Guttman's λ6 (if item dropped)
Greatest lower bound (if item dropped)
If item dropped plot
Order items
Order items by mean
Order items by KL-divergence
Order items by KS-distance
Item-rest correlation
Mean
Standard deviation
Plot Posteriors
Fix range to 0-1
Display Priors
Probability for: 0.7 < Reliability < 1
Shade posterior region in plot
Results
Bayesian Unidimensional Reliability
Bayesian Scale Reliability Statistics
Estimate McDonald's ω Cronbach's α
Posterior mean 0.774 0.786
95% CI lower bound 0.690 0.708
95% CI upper bound 0.852 0.856
Probability that Reliability Statistic is Larger than 0.70 and Smaller than 1.00
Probability
Statistic Prior Posterior
McDonald's ω 0.136 0.948
Cronbach's α 0.243 0.975
Posterior Plots
McDonald's ω
Cronbach's α
Density
0.69 0.852
0.708 0.856
McDonald's ω
Cronbach's α

Tutorial



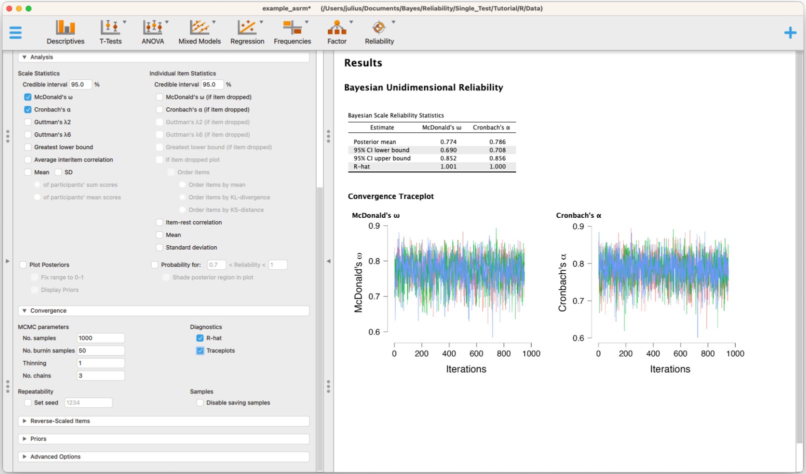
Bayesian Scale Reliability Statistics

Estimate	McDonald's ω	Cronbach's α
Posterior mean	0.774	0.786
95% CI lower bound	0.690	0.708
95% CI upper bound	0.852	0.856

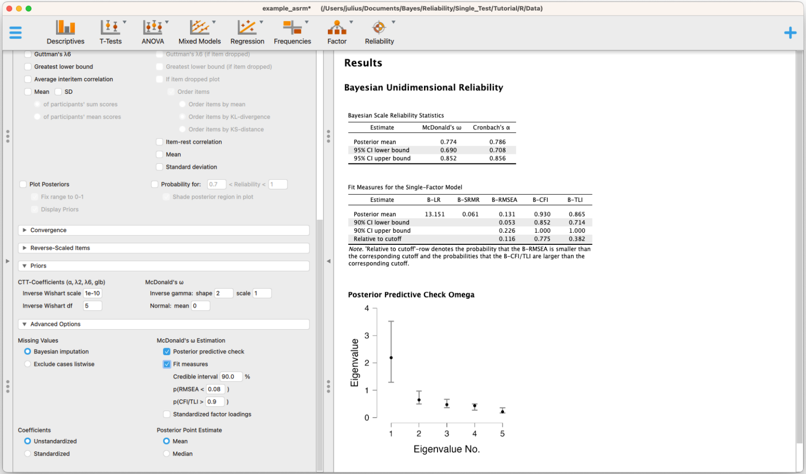
Bayesian Individual Item Reliability Statistics

	McDonald's ω (if item dropped)			Cronbach's α (if item dropped)		
Item	Posterior mean	Lower 95% CI	Upper 95% CI	Posterior mean	Lower 95% CI	Upper 95% CI
asrm_1	0.701	0.584	0.797	0.718	0.618	0.824
asrm_2	0.711	0.604	0.811	0.727	0.627	0.822
asrm_3	0.776	0.696	0.853	0.789	0.714	0.859
asrm_4	0.749	0.661	0.835	0.760	0.669	0.843
asrm_5	0.714	0.611	0.817	0.728	0.617	0.820

Tutorial


example_asrm* (/Users/julius/Documents/Bayes/Reliability/Single_Test/Tutorial/R/Data)
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Reliability
Analysis
Scale Statistics
Credible interval 95.0 %
McDonald's ω
Cronbach's α
Guttman's λ2
Guttman's λ6
Greatest lower bound
Average interitem correlation
Mean
SD
of participants' sum scores
of participants' mean scores
Individual Item Statistics
Credible interval 95.0 %
McDonald's ω (if item dropped)
Cronbach's α (if item dropped)
Guttman's λ2 (if item dropped)
Guttman's λ6 (if item dropped)
Greatest lower bound (if item dropped)
If item dropped plot
Order items
Order items by mean
Order items by KL-divergence
Order items by KS-distance
Item-rest correlation
Mean
Standard deviation
Plot Posteriors
Fix range to 0-1
Display Priors
Probability for: 0.7 < Reliability < 1
Shade posterior region in plot
Convergence
MCMC parameters
No. samples 1000
No. burnin samples 50
Thinning 1
No. chains 3
Diagnostics
R-hat
Traceplots
Repeatability
Set seed 1234
Samples
Disable saving samples
Reverse-Scaled Items
Priors
Advanced Options
Results
Bayesian Unidimensional Reliability
Bayesian Scale Reliability Statistics
Estimate | McDonald's ω | Cronbach's α
Posterior mean | 0.774 | 0.786
95% CI lower bound | 0.690 | 0.708
95% CI upper bound | 0.852 | 0.856
R-hat | 1.001 | 1.000
Convergence Traceplot
McDonald's ω
Cronbach's α
0.9
0.8
0.7
0.6
0 200 400 600 800 1000
Iterations

Tutorial


example_asrm* (/Users/julius/Documents/Bayes/Reliability/Single_Test/Tutorial/R/Data)
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Reliability
Guttman's λ6
Greatest lower bound
Average interitem correlation
Mean
SD
of participants' sum scores
of participants' mean scores
Guttman's λ6 (if item dropped)
Greatest lower bound (if item dropped)
If item dropped plot
Order items
Order items by mean
Order items by KL-divergence
Order items by KS-distance
Item-rest correlation
Mean
Standard deviation
Plot Posteriors
Fix range to 0-1
Display Priors
Probability for: 0.7 < Reliability < 1
Shade posterior region in plot
Convergence
Reverse-Scaled Items
Priors
CTT-Coefficients (α, λ2, λ6, glb)
Inverse Wishart scale 1e-10
Inverse Wishart df 5
McDonald's ω
Inverse gamma: shape 2 scale 1
Normal: mean 0
Advanced Options
Missing Values
Bayesian imputation
Exclude cases listwise
McDonald's ω Estimation
Posterior predictive check
Fit measures
Credible interval 90.0 %
p(RMSEA < 0.08)
p(CFI/TLI > 0.9)
Standardized factor loadings
Coefficients
Unstandardized
Standardized
Posterior Point Estimate
Mean
Median
Results
Bayesian Unidimensional Reliability
Bayesian Scale Reliability Statistics
Estimate | McDonald's ω | Cronbach's α
Posterior mean | 0.774 | 0.786
95% CI lower bound | 0.690 | 0.708
95% CI upper bound | 0.852 | 0.856
Fit Measures for the Single-Factor Model
Estimate | B-LR | B-SRMR | B-RMSEA | B-CFI | B-TLI
Posterior mean | 13.151 | 0.061 | 0.131 | 0.930 | 0.865
90% CI lower bound | | | 0.053 | 0.852 | 0.714
90% CI upper bound | | | 0.226 | 1.000 | 1.000
Relative to cutoff | | | 0.116 | 0.775 | 0.382
Note. 'Relative to cutoff'-row denotes the probability that the B-RMSEA is smaller than the corresponding cutoff and the probabilities that the B-CFI/TLI are larger than the corresponding cutoff.
Posterior Predictive Check Omega
Eigenvalue
Eigenvalue No.
0
1
2
3
4
5

Outline

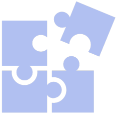
Conclusions

Part I – Psychometric models:

- Lower bounds remain useful under certain conditions
- FA-reliability is different from CTT-reliability
- Coefficient α is a lower bound to the reliability as defined by CTT

Part II – Uncertainty estimation:

- Uncertainty estimation is imperative in reliability analysis
- The posterior distribution of reliability coefficients is highly practical
- R-package and JASP implementation help researchers change their reliability reporting routine



Thank you for your attention!

Appendix

CTT-Coefficients (α, λ_2, glb)

Calculated from the data covariance matrix, $\mathbf{\Sigma}$:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\text{tr}(\mathbf{\Sigma})}{\mathbf{\Sigma}}\right) \tag{8}$$

$$\lambda_2 = \frac{\mathbf{\Sigma} - \text{tr}(\mathbf{\Sigma}) + \sqrt{\frac{k}{k-1}\, c}}{\mathbf{\Sigma}} \tag{9}$$

$$\text{glb} = 1 - \frac{\text{tr}(\mathbf{\Sigma}_E)}{\mathbf{\Sigma}} \tag{10}$$

FA-Coefficients

- Unidimensional data → based on single-factor model:

$$\omega_u = \frac{(\sum \lambda)^2}{(\sum \lambda)^2 + \sum \psi} \tag{11}$$

- Multidimensional data → based on bi-factor model:

$$\omega_t = \frac{\sum \Lambda^2}{\sum \Lambda^2 + \sum \psi} \tag{12}$$

$$\omega_h = \frac{(\sum \lambda_g)^2}{(\sum \lambda_g)^2 + \sum \psi}. \tag{13}$$

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

Coefficient α Rejoinder

Article II:

Sijtsma, K., & Pfadt, J. M. (2021b). Rejoinder: The future of reliability. *Psychometrika*, *86*(4), 887–892. https://doi.org/10.1007/s11336-021-09807-9

- Rejoinder to comments by Bentler, Ellis, and Cho
- Sound psychological theory should be at the core of any measurement
- The theory informs the measurement model which informs the reliability approach
- Disentangling target and non-target influences is not validity research
- In relation to reliability two main research areas are often overlooked:
  - How does reliability relate to the power of statistical tests?
  - How to properly indicate the measurement error of an individual?

Studies to investigate the performance of reliability coefficients use narrow data generation schemes → How do the coefficients perform with a wide range of data structures?

Article III

Pfadt, J. M., & Sijtsma, K. (2022). Statistical properties of lower bounds and factor analysis methods for reliability estimation. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology: The 86th Annual Meeting of the Psychometric Society, virtual, 2021* (pp. 51–63). Springer International Publishing. https://doi.org/10.1007/978-3-031-04572-1_5

Simulation Study

- Uni- and Multidimensional data generated from IRT models (conceptually closer to CTT), and an FA models
- Coefficients: α, λ_2, λ_4, glb, ω_u, ω_h, ω_t
- Misspecification condition:
  - Case (1):
    - Population model is multidimensional with a common factor
    - Researcher assumes unidimensionality $\rightarrow$ coefficient ω_u
  - Case (2):
    - Population model is purely multidimensional with no common factor
    - Researcher assumes a common factor $\rightarrow$ estimates coefficients α, λ_2, λ_4, glb, ω_h, ω_t

Results – Unidimensional Data

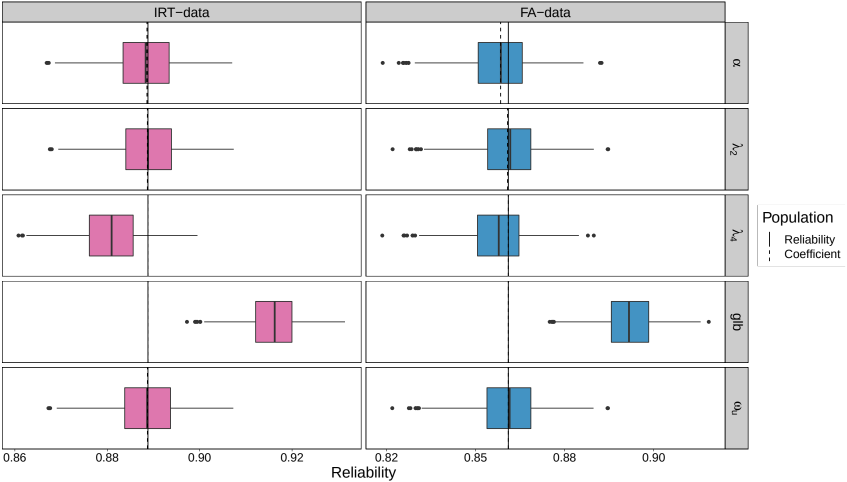


Figure 1. The point estimates of the coefficients across 1,000 simulation runs for $k = 18$ items and sample size of $n = 500$. In the IRT-conditions the data were generated from a 2-parameter graded response model. In the FA-conditions the data were generated from a single-factor model.

Results: Multidimensional Data

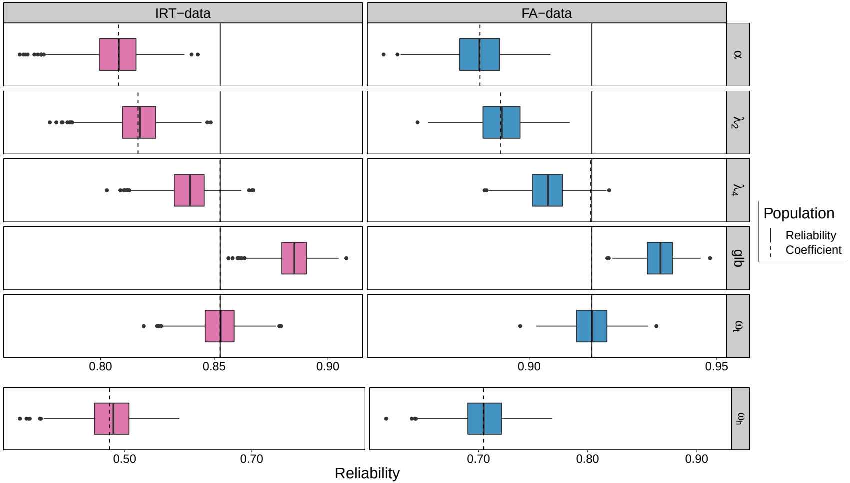


Figure 2. The point estimates of the coefficients across 1,000 simulation runs for $k = 18$ items and sample size of $n = 500$. In the IRT-conditions the data were generated from a 2-parameter graded response model with three latent variables and intercorrelations of .3. In the FA-conditions the data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

Results: Misspecified Models

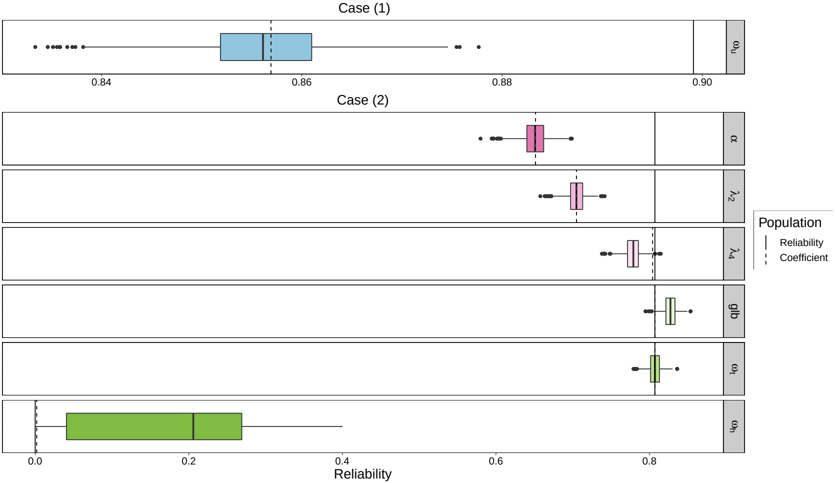


Figure 3. The point estimates of the coefficients across 1,000 simulation runs with $n = 1,000$. The data for Case (1) was generated from a second-order factor model with three primary latent variables. The data for Case (2) was generated from a factor model with three latent variables and no intercorrelations.

Simulation Study

Results summary:

- No meaningful differences between the IRT and FA conditions
- With unidimensional data, most coefficients performed well
- With multidimensional data the ω-coefficients performed well

Conclusions:

- When data are unidimensional the choice of a reliability coefficient is virtually arbitrary
- When data are multidimensional use an FA-coefficient
- When using an FA-coefficient confirm model fit

Simulation Study – Bayesian Single Test Reliability

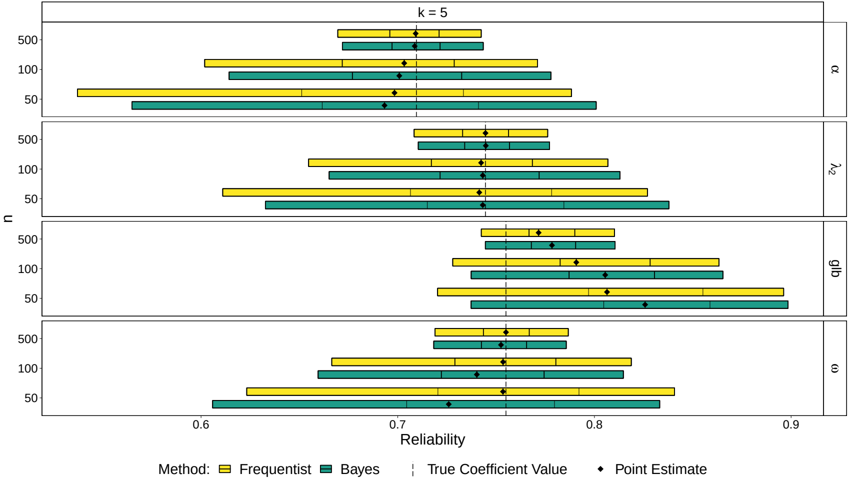


Figure 5. Simulation results for the medium-correlation condition with $k = 5$ items. The endpoints of the bars are the mean 95% uncertainty interval limits. The 25%- and 75%-quartiles are indicated with vertical line segments.

Simulation Study – Bayesian Single Test Reliability

Results summary:

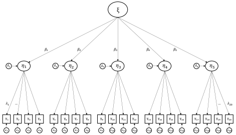
- The credible intervals for coefficients α, λ_2, and ω_u performed satisfactory,
- The Bayesian point estimation was slightly worse than the classical (frequentist) in small samples
- The results for the classical bootstrap confidence intervals and the Bayesian credible intervals generally agreed

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

Introduction – Bayesian Multidimensional Reliability

- Coefficients ω_t for the total reliability and ω_h for the g-factor reliability (see Equations 12 and 13)
- 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

→ Develop Bayesian versions of ω_t and ω_h

→ 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

Results summary:

- Out of the confidence intervals, the BCA, 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, otherwise the BCA interval

References I

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press. https://doi.org/10.1201/b16018

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930–944. https://doi.org/10.1177/0013164406288165

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(2), 251–270. https://doi.org/10.1207/S15328007SEM0702_6

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

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.

References II

Lucke, J. F. (2005). "Rassling the hog": The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied Psychological Measurement*, *29*(2), 106–125. https://doi.org/10.1177/0146621604272739

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*, *227*(3), 186–194. https://doi.org/10.1027/2151-2604/a000377

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. (2022). Statistical properties of lower bounds and factor analysis methods for reliability estimation. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology: The 86th Annual Meeting of the Psychometric Society, virtual, 2021* (pp. 51–63). Springer International Publishing. https://doi.org/10.1007/978-3-031-04572-1_5

Pfadt, J. M., van den Bergh, D., & Moshagen, M. (in press). Classical and Bayesian uncertainty intervals for the reliability of multidimensional scales. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2022.2124162

Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E.-J. (2022). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*, *57*(4), 620–641. https://doi.org/10.1080/00273171.2021.1891855

References III

Pfadt, J. M., van den Bergh, D., Sijtsma, K., & Wagenmakers, E.-J. (in press). A tutorial on Bayesian single-test reliability analysis with JASP. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01778-0

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. (2021a). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*(4), 843–860. https://doi.org/10.1007/s11336-021-09789-8

Sijtsma, K., & Pfadt, J. M. (2021b). Rejoinder: The future of reliability. *Psychometrika*, *86*(4), 887–892. https://doi.org/10.1007/s11336-021-09807-9

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher*, *22*(2), 209–213. https://doi.org/10.1007/s40299-013-0075-z