Simulation Study Preregistration for The Interval Truth Model: A Consensus Model for Continuous Bounded Interval Responses

Matthias Kloft, Björn S. Siepe, Daniel W. Heck

June 26, 2024

Version: 0.1.0
Last updated: 2023-10-31
This preregistration is based on the preregistration template designed by Björn S. Siepe, František Bartoš, Tim P. Morris, Anne-Laure Boulesteix, Daniel W. Heck, and Samuel Pawel
Retrieved from: https://www.overleaf.com/latex/templates/ademp-prereg-simulation-study-template/dkhtxjtmpbfj

1 Instructions

2 General Information

2.1 What is the title of the project?

Answer: The Interval Truth Model: A Consensus Model for Continuous Bounded Interval Responses

2.2 Who are the current and future project contributors?

Answer: Matthias Kloft, Björn S. Siepe, Daniel W. Heck

2.3 Provide a description of the project.

Explanation: This can also include empirical examples that will be analyzed within the same project, especially if the analysis depends on the results of the simulation.
Answer: We develop a consensus model that applies to interval responses on continuous bounded response scales. The model builds on the work of Anders et al. (2014), Mayer and Heck (2023), and Smithson and Broomell (2024).

2.4 Did any of the contributors already conduct related simulation studies on this specific question?

Explanation: This includes preliminary simulations in the context of the current project.
Answer: We performed preliminary simulations with low numbers of iterations $n < 20$ and a single data-generating process to check our simulation code for errors.

3 Aims

3.1 What is the aim of the simulation study?

Explanation: The aim of a simulation study refers to the goal of the research and shapes subsequent choices. Aims are typically related to evaluating the properties of a method (or multiple methods) with respect to a particular statistical task. Possible tasks include 'estimation', 'hypothesis testing', 'model selection', 'prediction', or 'design'. If possible, try to be specific and not merely state that the aim is to 'investigate the performance of method X under different circumstances'.
Answer: The simulation study aims to explore the estimation performance of the model with respect to bias and mean-squared error in realistic scenarios of use.

4 Data-Generating Mechanism

4.1 How will the parameters for the data-generating mechanism (DGM) be specified?

Explanation: Answers include 'parametric based on real data', 'parametric', or 'resampled'. Parametric based on real data usually refers to fitting a model to real data and using the parameters of that model to simulate new data. Parametric refers to generating data from a known model or distribution, which may be specified based on theoretical or statistical knowledge, intuition, or to test extreme values. Resampled refers to resampling data from a certain data set, in which case the true data-generating mechanism is unknown. The answer to this question may include an explanation of from which distributions (with which parameters) values are drawn, or code used to generate parameter values. If the DGM parameters are based on real data, please provide information on the data set they are based on and the model used to obtain the parameters. Also, indicate if any of the authors are already familiar with the data set, e.g., analyzed (a subset of) it.
Answer: We will simulate data based on a parametric data-generating mechanism, which we explain in more detail below.

4.2 What will be the different factors of the data-generating mechanism?

Explanation: A factor can be a parameter/setting/process/etc. that determines the data-generating mechanism and is varied across simulation conditions.
Answer: We will vary the following factors:

- Number of respondents: $\{10, 50, 100, 200\}$
- Number of items: $\{5, 10, 20, 40\}$

These numbers are selected to cover a range of combinations from scenarios where there are few items and also only a few experts (e.g., a company has ten expert employees rate the risk of a security breach for five departments) to scenarios where a lot of raters and items might be available (e.g., a forecasting challenge).

4.3 If possible, provide specific factor values for the DGM as well as additional simulation settings.

Explanation: This may include a justification of the chosen values and settings.
Answer: We will use the following data-generating mechanism for the interval re-

sponse $\boldsymbol{Z}_{ij}$ of respondent i on item j on an unbounded scale:

$$\begin{aligned}
&\boldsymbol{Z}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}),\\
&\boldsymbol{\mu}_{ij} = \left[T_j^{loc} a_i^{loc} + b_i^{loc},\; T_j^{wid} + b_i^{wid}\right]\\
&\boldsymbol{\Sigma}_{ij} = \begin{bmatrix} \left(a_i^{loc} \frac{1}{E_i^{loc}} \frac{1}{\lambda_j^{loc}}\right)^2 & 0 \\ 0 & \left(\frac{1}{E_i^{wid}} \frac{1}{\lambda_j^{wid}}\right)^2 \end{bmatrix}
\end{aligned} \tag{1}$$

The parameter interpretations are defined below, together with the respective values that will be used to simulate the data. To arrive at the actual interval response in the bounded simplex space, we first will transform the unbounded interval response $\boldsymbol{Z}_{ij}$ using the inverse isometric log-ratio function (see Smithson & Broomell, 2024):

$$\begin{aligned}
\boldsymbol{Y} &= \left[\frac{exp\Big(\sqrt{2}\, Z_{ij1}\Big)}{\Sigma},\; \frac{exp\left(\sqrt{\frac{3}{2}}\, Z_{ij2} + \frac{Z_{ij1}}{\sqrt{2}}\right)}{\Sigma},\; \frac{1}{\Sigma}\right]^\top\\
\text{with} \quad \Sigma &= exp\Big(\sqrt{2} + Z_{ij1}\Big) + exp\left(\sqrt{\frac{3}{2}}\, Z_{ij2} + \frac{Z_{ij1}}{\sqrt{2}}\right) + 1
\end{aligned} \tag{2}$$

For model estimation, the data is transformed back to the unbounded space by the isometric log-ratio function (see Smithson & Broomell, 2024):

$$\boldsymbol{Z}_{ij} = \left[\sqrt{\frac{1}{2}} \log\left(\frac{Y_{ij1}}{Y_{ij3}}\right),\; \sqrt{\frac{2}{3}} \left(\log \frac{Y_{ij2}}{\sqrt{Y_{ij1} Y_{ij3}}}\right)\right]^\top \tag{3}$$

This back-and-forth transformation is a redundant step in our simulation since we use the same transformation in the data generation as in the model fitting. However, it matters if one uses different transformation functions for each step (see Section 8.4). The data-generating model will be explained in detail in the article for this simulation study and is presented here for completeness.

In the following, we will justify our choices of hyperparameter values. Overall, our aim is to generate plausible distributions of response intervals. We derive the hyperparameters from response intervals representing conditions of distributional modes or boundaries. First, we determine the means and standard deviations of latent true interval locations and widths by propagating the interval $[.40, .60]$ through the isometric log-ratio transformation function (Smithson & Broomell, 2024), which results in $\mu_{T^{loc}} = 0$ and $\mu_{T^{wid}} = -0.57$. In the same way, we derive the standard deviation for the latent true interval locations from the interval $[.98, .99]$, which yields an extreme location as a boundary value. By dividing the distance of this boundary value from the mean $\mu_{T^{loc}}$ by four we arrive at $\sigma_{T^{loc}} = 0.81$. Similarly, we derive the standard deviation for the latent true widths from the interval $[.495, .505]$, which yields $\sigma_{T^{wid}} = 0.66$. The means of the hyperparameters that are fixed to zero or one in the model (λ, a, b) are fixed to the same values in the data-generating process. The means for the inverted person proficiency parameters $\frac{1}{E}$ defined on the log-scale are scaled to the standard deviation of the respective T parameters, e.g., $\mu_{E^{wid}} = log(\sigma_{T^{wid}}) = log(0.66)$. The hyperparameter standard deviations for the parameters other than T are derived in two ways: First, for the standard deviations of shifting biases b, we divide the standard deviation of the respective dimension of true intervals T by three. For example, $\sigma_{b^{wid}} = \sigma_{T^{wid}}/3 = 0.65/3 = 0.22$. The standard deviations for the other parameters on the log scale are fixed to 0.3. The person proficiencies (E^{loc}, E^{wid}) and item discernibilities ($\lambda^{loc}, \lambda^{wid}$) are inverted after sampling, i.e. multiplied by -1 on the log scale such as $-log(\lambda^{loc})$. This means that the parameters are sampled using a variance parameterization but implemented in the model using a precision parameterization. This gives us the advantage of comparability regarding the variances of the other parameters, especially $\sigma_{T^{loc}}, \sigma_{T^{wid}}$. These settings for the precision parameters (E, λ) ensure that the resulting parameter values on the positive scale lie mostly in the range of plausible values from 0.3 to 2. We draw the actual true parameters from normal distributions using the described means and standard deviations as hyperparameters. Normal distributions can be used since all parameters are defined on the latent (unbounded) scale:

- For each item:
  - True interval location: $T^{loc} \sim N(0, 0.81)$
  - True interval width: $T^{wid} \sim N(-0.57, 0.65)$
  - Discernibility for location: $-\log(\lambda^{loc}) \sim \mathcal{N}(0, 0.3)$
  - Discernibility for width: $-\log(\lambda^{wid}) \sim \mathcal{N}(0, 0.3)$
- For each respondent:
  - Proficiency for location: $-\log(E^{loc}) \sim \mathcal{N}(log(0.81), 0.3)$
  - Proficiency for width: $-\log(E^{wid}) \sim \mathcal{N}(log(0.65), 0.3)$
  - Scaling bias for location: $\log(a^{loc}) \sim \mathcal{N}(0, 0.3)$
  - Shifting bias for location: $b^{loc} \sim \mathcal{N}(0, 0.27)$
  - Shifting bias for width: $b^{wid} \sim \mathcal{N}(0, 0.22)$

4.4 If there is more than one factor: How will the factor levels be combined and how many simulation conditions will this create?

Explanation: Answers include ‘fully factorial’, ‘partially factorial’, ‘one-at-a-time’, or ‘scattershot’. Fully factorial designs are designs in which all possible factor combinations are considered. Partially factorial designs denote designs in which only a subset of all possible factor combinations are used. One-at-a-time designs are designs where each factor is varied while the others are kept fixed at a certain value. Scattershot designs include distinct scenarios, for example, based on parameter values from real-world data.

Answer: We will vary the conditions in a fully factorial manner. This will result in 4 (number of respondents) × 4 (number of items) = 16 simulation conditions.

5 Estimands and Targets

5.1 What will be the estimands and/or targets of the simulation study?

Explanation: Please also specify if some targets are considered more important than others, i.e., if the simulation study will have primary and secondary outcomes.
Answer: The simulation study's main targets will be the latent true interval locations and widths. We will also compare these model estimates to simple means for each item, i.e., averaging after applying the isometric log-ratio transformation instead of estimating the parameters of the consensus model. Further, we will investigate the recovery of the other person and item parameters.

6 Methods

6.1 How many and which methods will be included and which quantities will be extracted?

Explanation: Be as specific as possible regarding the methods that will be compared, and provide a justification for both the choice of methods and their model parameters. This can also include code which will be used to estimate the different methods or models in the simulation with all relevant model parameters. Setting different prior hyperparameters might also be regarded as using different methods. Where package defaults are used, state this. Where they are not used, state what values are used instead.
Answer: We keep the description of methods short here for reasons of brevity. More details on their background will be provided in the manuscript. We will estimate the same model for all generated data sets in a Bayesian framework using `Stan` (Stan Development Team, 2023) in `R` (R Core Team, 2023) via `cmdstanr` (Gabry et al., 2023).

7 Performance Measures

7.1 Which performance measures will be used?

Explanation: Please provide details on why they were chosen and on how these measures will be calculated. Ideally, provide formulas for the performance measures to avoid ambiguity. Some models in psychology, such as item response theory or time series models, often contain multiple parameters of interest, and their number may vary across conditions. With a large number of estimated parameters, their performance measures are often combined. If multiple estimates are aggregated, specify how this aggregation will be performed. For example, if there are multiple parameters in a particular condition, the mean of the individual biases of these parameters or the bias of each individual parameter may be reported.

Answer: Our primary performance measure is the absolute bias of the latent true interval location and width (T_j^{loc}, T_j^{wid}), which we define as follows:

$$\widehat{\text{AbsBias}} = \frac{\sum_{i=1}^{n_{\text{sim}}} \sum_{p=1}^{n_{\text{ind}}} 0.5\big(|\hat{T}_{pi}^{loc} - T^{loc}| + |\hat{T}_{pi}^{wid} - T^{wid}|\big)}{n_{\text{sim}} \times n_{\text{ind}}},$$

where n_{ind} is the number of participants in a specific condition and n_{sim} is the number of replications, which is the same for each condition.

We compute the mean of the (absolute) bias of the location and width because we expect that there may be some compensatory behavior of the estimates. We will additionally output the individual biases and visualize them jointly in a scatter plot to assess potential compensatory behavior.

We will additionally calculate the mean squared error (MSE) for the bivariate vector of true intervals, $([T_j^{loc}, T_j^{wid}])$,

$$\widehat{\text{MSE}} = \frac{\sum_{i=1}^{n_{\text{sim}}} \sum_{p=1}^{n_{\text{ind}}} 0.5\Big((\hat{T}_{pi}^{loc} - T^{loc})^2 + (\hat{T}_{pi}^{wid} - T^{wid})^2\Big)}{n_{\text{sim}} \times n_{\text{ind}}}$$

We will also calculate the MSE for the location and width individually. We estimate the MCSE of these performance measures via bootstrapping.

We will additionally report the number of non-converged/missing replications per method per condition. We will use the built-in features of the `SimDesign` package (Chalmers & Adkins, 2020) to handle non-convergence. Regarding the used MCMC sampler, we will also track the number of divergent transitions per condition as well as the $\hat{R}$ statistic.

7.2 How will Monte Carlo uncertainty of the estimated performance measures be calculated and reported?

Explanation: Ideally, Monte Carlo uncertainty can be reported in the form of Monte Carlo Standard Errors (MCSEs). Please see Siepe et al. (2023) and Morris et al. (2019) for a list of formulae to calculate the MCSE related to common performance measures, more accurate jackknife-based MCSEs are available through the `rsimsum` (Gasparini, 2018) and `simhelpers` (Joshi & Pustejovsky, 2022) R packages, the `SimDesign` (Chalmers & Adkins, 2020) R package can compute confidence intervals for performance measures via bootstrapping. Monte Carlo uncertainty can additionally be visualized using plots appropriate for illustrating variability, such as MCSE error bars, histograms, box plots, or violin plots of performance measure estimates, if possible (e.g., bias).

Answer: We will use a bootstrapping procedure to calculate MCSEs (see our answer to the last question) and report them in plots and/or tables.

7.3 How many simulation repetitions will be used for each condition?

Explanation: Please also indicate whether the chosen number of simulation repetitions is based on sample size calculations, on computational constraints, rules of thumb, *or any other heuristic or combination of these strategies. Formulas for sample size planning in simulation studies are provided in Siepe et al. (2023). If there is a lack of knowledge on a quantity for computing the Monte Carlo standard error (MCSE) of an estimated performance measure (e.g., the variance of the estimator is needed to compute the MCSE for the bias), pilot simulations may be needed to obtain a guess for realistic/worst-case values.*

Answer:

Due to the large computational demand of our simulation study, we decided on our sample size as follows: We aimed for an MCSE of $\leq .05$ for our primary performance measure (the absolute bias for the latent true interval location and width) in all conditions. We deemed 1000 repetitions computationally reasonable. After 1000 repetitions, we will check the MCSEs in all conditions. Should they not fulfill the criterion above, we will incrementally add repetitions in steps of 250 until they do.

7.4 How will missing values due to non-convergence or other reasons be handled?

Explanation: 'Convergence' means that a method successfully produces the outcomes of interest (e.g., an estimate, a prediction, a p-value, a sample size, etc.) that are required for estimating the performance measures. Non-convergence of some iterations or whole conditions of simulation studies occurs regularly, e.g., for numerical reasons. It is possible to impute non-converged iterations, exclude all non-converged iterations, or implement mechanisms that repeat certain parts of the simulation (such as data generation or model fitting) until convergence is achieved. Further, it is important to consider at which proportion of failed iterations a whole condition will be excluded from the analysis.

Answer: If we observe any non-convergence, we exclude the non-converged cases and report the number of non-converged cases per method and condition. All performance measures will be based on the converged cases only.

7.5 How do you plan on interpreting the performance measures? (optional)

Explanation: It can be specified what a 'relevant difference' in performance, or what 'acceptable' and 'unacceptable' levels of performance might be to avoid post-hoc interpretation of performance. Furthermore, some researchers use regression models to analyze the results of simulations and compute effect sizes for different factors, or to assess the strength of evidence for the influence of a certain factor (Chipman & Bingham, 2022; Skrondal, 2000). If such an approach will be used, please provide as many details as possible on the planned analyses.

Answer: We compare the model estimates of the latent true intervals against simple means as a simpler competitor model. Given that the data are generated from our model, we expect the model estimates to perform better than the simple means. If that is not the case, the added complexity of our model may not be worth the effort compared to alternative aggregation strategies. We further expect that higher numbers of respondents will lead to better performance of item parameters, and, vice versa, that higher numbers of items will lead to better performance of person parameters.

8 Other

8.1 Which statistical software/packages do you plan to use?

Explanation: Likely, not all software used can be prespecified before conducting the simulation. However, the main packages used for model fitting are usually known in advance and can be listed here, ideally with version numbers.
Answer: We will use the following packages for `R` version 4.4.1 (R Core Team, 2023) in their most recent versions at the time of writing: `SimDesign` (Chalmers & Adkins, 2020) for setting up and conducting the simulation study, `cmdstanr` (Gabry et al., 2023) as the R interface to Stan (Stan Development Team, 2023), and the `posterior` (Bürkner et al., 2023) and `bayesplot` package (Gabry & Mahr, 2024) for handling and visualizing MCMC output. Additional packages used for data wrangling and minor tasks will be provided in the supplementary materials of our main manuscript.

8.2 Which computational environment do you plan to use?

Explanation: Please specify the operating system and its version which you intend to use. If the study is performed on multiple machines or servers, provide information for each one of them, if possible.
Answer: We will run the simulation study on a Linux machine with an Ubuntu 22.04.4 LTS distribution. The complete output of `sessionInfo()` will be saved and reported in the supplementary materials.

8.3 Which other steps will you undertake to make simulation results reproducible? (optional)

Explanation: This can include sharing the code and full or intermediate results of the simulation in an open online repository. Additionally, this may include supplemental materials or interactive data visualizations, such as a shiny application.
Answer: We will upload the fully reproducible simulation script and a data set containing all relevant summary measures of all simulation studies to OSF.

8.4 Is there anything else you want to preregister? (optional)

Explanation: For example, the answer could include the most likely obstacles in the simulation design, and the plans to overcome them.
Answer: We will conduct a preliminary, smaller simulation study to test two alternative link functions for modeling interval responses. The setup described above uses the isometric log-ratio function described in Smithson and Broomell (2024). These authors also propose an alternative transformation based on a stick-breaking procedure,

which uses an amalgamation log-ratio balance. We want to compare this alternative transformation to the isometric log-ratio transformation. We will generate and fit data using both transformations as link functions in a fully crossed design. In other words, each link function will be used on data simulated with itself and with the other link function. For the alternative link function, we make a slight adjustment to the hyperparameters: We determine the mean of the latent true interval by propagating the interval $[.425, .575]$ through the alternative transformation function, which results in $\mu_{T^{loc}} = 0$ and $\mu_{T^{wid}} = -1.73$. Otherwise, the data-generating mechanism would have produced an implausibly high amount of broad interval widths $> .5$ on the bounded scale. The rest of the hyperparameters are derived in the same way as described in Section 4.3. To keep this preliminary study of a small detail of our modeling strategy as simple as possible, we will only use one combination of 200 respondents and 30 items. We will compare the performance of the models using different link functions with the same performance measures as in the main simulation study described above. We plan to use the isometric log-ratio transformation for the main simulation study because it scales the interval width more strongly according to its location. Also, Kloft and Heck (2024) found a better model fit to empirical data compared to the alternative transformation. However, should we find that the alternative transformation function performs comparatively better, we will use that transformation in the main simulation study.

We will additionally re-analyze previously collected data from Ellerby et al. (2022) and Kloft and Heck (2024).

References

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13. https://doi.org/10.1016/j.jmp.2014.06.001

Bürkner, P.-C., Gabry, J., Kay, M., & Vehtari, A. (2023). *Posterior: Tools for working with posterior distributions* (Version 1.5.0) [Software R package]. https://mc-stan.org/posterior/

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chipman, H., & Bingham, D. (2022). Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments. *Canadian Journal of Statistics*, *50*(4), 1228–1249. https://doi.org/10.1002/cjs.11719

Ellerby, Z., Wagner, C., & Broomell, S. B. (2022). Capturing richer information: On establishing the validity of an interval-valued survey response mode. *Behavior Research Methods*, *54*(3), 1240–1262. https://doi.org/10.3758/s13428-021-01635-0

Gabry, J., Češnovar, R., & Johnson, A. (2023). *Cmdstanr: R interface to 'cmdstan'* (Version 0.7.0) [Software R package]. https://mc-stan.org/cmdstanr/

Gabry, J., & Mahr, T. (2024). *Bayesplot: Plotting for bayesian models* (Version 1.11.1) [Software R package]. https://mc-stan.org/bayesplot/

Gasparini, A. (2018). Rsimsum: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, *3*(26), 739. https://doi.org/10.21105/joss.00739

Joshi, M., & Pustejovsky, J. (2022). *Simhelpers: Helper functions for simulation studies* [R package version 0.1.2]. https://CRAN.R-project.org/package=simhelpers

Kloft, M., & Heck, D. W. (2024). *Discriminant validity of interval responses: Investigating the dimensional structure of interval response widths using a novel multivariate-logit transformation*. PsyArXiv. https://doi.org/https://doi.org/10.31234/osf.io/esvxk

Mayer, M., & Heck, D. W. (2023). Cultural consensus theory for two-dimensional location judgments. *Journal of Mathematical Psychology*, *113*, 102742. https://doi.org/10.1016/j.jmp.2022.102742

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.4.1) [Software]. Vienna, Austria. https://www.R-project.org/

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D., & Pawel, S. (2023). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting [Preprint]. https://doi.org/10.31234/osf.io/ufgy6

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167. https://doi.org/10.1207/s15327906mbr3502_1

Smithson, M., & Broomell, S. B. (2024). Compositional data analysis tutorial: Psychological Methods. *Psychological Methods*, *29*(2), 362–378. https://doi.org/10.1037/met0000464

Stan Development Team. (2023). *Stan Modeling Language Users Guide and Reference Manual* (Version 2.33) [Software]. https://mc-stan.org