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

Matthias Kloft

University of Marburg

Björn S. Siepe

University of Marburg

Daniel W. Heck

University of Marburg

Author Note

Matthias Kloft https://orcid.org/0000-0003-1845-6957

Björn S. Siepe https://orcid.org/0000-0002-9558-4648

Daniel W. Heck https://orcid.org/0000-0002-6302-9252

Correspondence. Correspondence concerning this article should be addressed to Matthias Kloft, Psychological Methods Lab, Department of Psychology, University of Marburg, Gutenbergstr. 18, 35032 Marburg. E-mail: kloft(hat)uni-marburg(bot)de

Author Contributions. ***MK***: conceptualization, formal analysis, methodology, investigation, data curation, writing — original draft, writing — review & editing. ***BSS***: conceptualization, investigation, formal analysis, methodology, investigation, writing — original draft, writing – review & editing. ***DW***: conceptualization, formal analysis, methodology, writing – review & editing, supervision.

Open Practices Statement. All data and analysis scripts as well as the preregistration of the simulation study are available at the Open Science Framework (OSF): https://osf.io/r32by/.

Funding Statement. This research received no specific grant funding from any funding agency, commercial, or not-for-profit sectors.

Competing Interests. The authors have no relevant financial or non-financial interests to disclose.

Ethical Standards. We reanalyzed empirical data which we had collected in a previous study (Kloft & Heck, 2024). This study was performed in line with the guidelines of the German Psychological Society (DGPs). The Department of Psychology at the University of Marburg grants an exemption from ethics approval for online studies such as ours, where there is no deception of participants and no sensitive topics or materials are presented to participants.

Abstract

Cultural Consensus Theory (CCT) leverages shared knowledge between individuals to optimally aggregate answers to questions for which the underlying truth is unknown. Existing CCT models have predominantly focused on unidimensional point truths using dichotomous, polytomous, or continuous response formats. However, certain domains such as risk assessment or interpretation of verbal quantifiers may require a consensus focused on intervals, capturing a range of relevant values. We introduce the Interval Truth Model (ITM), a novel extension of CCT designed to estimate consensus intervals from continuous bounded interval responses. We use a Bayesian hierarchical modeling approach to estimate latent consensus intervals. In a simulation study, we show that, under the conditions studied, the ITM performs better than using simple means of the responses. We then apply the model to empirical judgments of verbal quantifiers.

Keywords: Continuous bounded responses, cultural consensus theory, interval responses, Bayesian modeling

1 Introduction

In psychological research, it is common practice to pose questions to respondents for which the correct answer is not known. This may be a forecast of the occurrence of some future event, for example “that same-sex marriage will be federally recognized by the end of Obama’s term (2017)” (Anders et al., 2014), where the correct answer can in principle be known or will reveal itself eventually. Correct answers are also unavailable in scenarios where the correct answer can change based on the context or the particular group of respondents. For example, one might be interested in judgments of affective valence regarding stimulus words like “accident” (Bradley & Lang, 1999) or in judgments of probabilities assigned to verbal quantifiers like “seldom” or “likely.” Such judgments can often be ambiguous and may systematically vary between groups or individuals, or even within a single individual, depending on the context in which the particular word is used (Karelitz & Budescu, 2004). In such scenarios, estimating the most valid answer from the given responses can be desirable.

Cultural consensus theory (CCT) was developed to solve this problem (Batchelder & Romney, 1988; Romney et al., 1986). It is based on the assumption that respondents belong to the same group or sub-population and share common knowledge about a particular knowledge domain, which is termed the *cultural consensus*. However, respondents may not all have the same level of expertise, and thus the quality of answers may vary among different respondents. The theory further assumes that weighting the responses by expertise will improve the overall accuracy of the aggregated judgments. CCT builds on these assumptions to estimate the cultural consensus by (a) aggregating all responses and (b) weighting each response by the inferred expertise of the respective respondent. To estimate the expertise of the respondents along with the cultural consensus, it is necessary to collect responses to multiple items in the same knowledge domain for each respondent. This can typically be done in a design in which respondents and items are fully crossed, but also in a non-fully crossed design. The consistency of a respondent’s answers across multiple items (relative to the answer patterns of other respondents) is then used to estimate their expertise in the respective domain.

Additionally, the discernibility of each item's cultural consensus is estimated across respondents and incorporated into the estimation of the cultural consensus.

Different consensus models for various combinations of response formats and modalities of the latent truth have been proposed. The initial consensus model, the general Condorcet model (Batchelder & Romney, 1988), used dichotomous responses to estimate binary truth values, e.g., for answers on a true-false general knowledge test. Following this, several model extensions have been proposed. The latent truth model (Batchelder & Anders, 2012) also accommodates dichotomous responses, but assumes that the true values of interest are continuous and lie between zero and one. For instance, respondents were asked for dichotomous judgments indicating whether a disease is contagious (Batchelder & Anders, 2012). While judgments about the perceived contagiousness of a disease can be assessed in a dichotomous response format, true contagiousness is more accurately represented in terms of probability, i.e., by a continuous value between zero and one. The latent truth values thus have a probabilistic meaning, while the observable responses are discrete, binary values of either zero or one. The continuous response model (Anders et al., 2014) extends this model to the case where responses are no longer dichotomous, but rather given on a continuous bounded response scale between zero and one. The model assumes that true values are continuous in a latent, unbounded space and are mapped onto the bounded response scale by a logit-link function. One application of this model concerns the forecasting of probabilities of future events, such as a large tsunami hitting the coast of a particular country (Anders et al., 2014). Anders et al. (2014) also incorporated a method for estimating multiple cultural consensuses for qualitatively different groups by combining CCT with latent class analysis. Another extension of the latent truth model, the latent truth rater model (Anders & Batchelder, 2015), maps continuous latent true values to categorical responses. An example application could be ratings of the grammatical acceptability of English phrases on a seven-point scale (Anders & Batchelder, 2015).

All models described above are unidimensional, as only one attribute is rated for each item. However, consensus models can also be applied to multidimensional ratings.

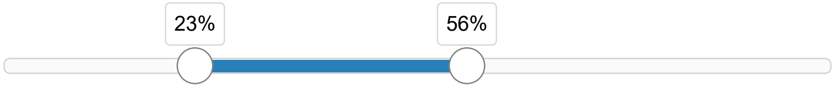
Mayer and Heck (2023) proposed a model for two-dimensional estimates of geographical locations on maps, where respondents had to estimate the location of cities such as London. In this case, both responses and latent true values refer to longitude and latitude and are thus continuous, two-dimensional vectors. In this specific example, the model assumes unbounded coordinates while actual locations are bounded due to geographic constraints such as oceans.

What all of the above models have in common is that they assume a single (uni- or two-dimensional) point as the latent, unknown truth for each item. However, in some domains, a point truth is too constraining and a range or interval of values may be more appropriate to represent a group's consensus. One example is the judgment of risks, for example in cyber-security (Ellerby et al., 2020). An interval rating in these cases can be conceptualized as an interval of risk estimates ranging from a best-case scenario to a worst-case scenario, i.e., a lower and an upper bound of the particular risk. The shared uncertainty of experts can be of interest to stakeholders and should be incorporated into these risk assessments. Another example concerns verbal quantifiers like "difficult" (Navarro et al., 2016) or "likely" (Karelitz & Budescu, 2004), which might be used to indicate how frequently or with which probability particular events such as extreme heat waves are happening (Harris et al., 2017). The use of such quantifiers is ambiguous, since there is no clear-cut convention in terms of numerical probabilities that should be assigned to particular quantifiers (except for words like "always" or "never"). An interval truth could represent a range of permissible probabilities that a particular word stands for in its pragmatic use.

Interval response formats such as the dual range slider (DRS) shown in Figure 1 may be a suitable solution for these types of applications. Sliders allow respondents to judge the lower and the upper bound of a range of values. Ellerby et al. (2022) found that respondents could adequately indicate the variability of different stimuli with an interval response format. In a multi-trait multi-method study, Kloft et al. (2024) found good test-retest reliability for both interval location and interval width. However, factor scores for interval widths did not show discriminant validity for the two personality scales

Figure 1

Dual Range Slider (DRS)



Note. Screenshot of the *noUiSlider* JavaScript range slider (Gersen, 2024) used in the empirical study (see Section 4). The scale ranges from 0% to 100%.

used (Extraversion and Conscientiousness). This finding was replicated in another study by Kloft and Heck (2024) in which the DRS response format was applied to different task domains. However, while the discriminant validity of the interval widths was low for the two personality scales, the factorial structure of the interval widths overall followed the structure of the different estimation tasks. These findings indicate that interval responses are suitable for estimation tasks in which some objectively quantifiable probability or frequency has to be rated. Although various methods for the aggregation of interval ratings have been proposed (Gaba et al., 2017; Lyon et al., 2015; Park & Budescu, 2015), a consensus model for this type of response format has not yet been developed. As a remedy, the present article aims at developing a consensus model that can be used to estimate weighted consensus intervals based on ratings collected via continuous bounded interval response formats like the DRS.

We focus on the case where the latent consensus itself is an interval. As discussed by Batchelder and Anders (2012) for unidimensional, dichotomous responses, different kinds of latent truths can be mapped onto the same response format with which observable ratings are collected. In the case of dichotomous responses, the latent truth can either be binary, i.e., true or false, or continuous, i.e., a probability between zero and one of being true or false. Similarly, in the case of collecting interval responses with the DRS response format on a scale from zero to one, the latent truth can be a single point in $[0, 1]$ such as the consensus probability of an event happening. However, the latent truth can also be a consensus interval in $[0, 1]$ if a range of values is permissible. For instance, in the example of verbal quantifiers, the word “often” could be associated with a consensus interval of $[.60, .80]$. Depending on the type of latent truth and the substantive application, different psychological constructs can be measured by interval responses (see also Kloft & Heck, 2024, for a discussion of relevant domains and psychological constructs). In the case of a point truth model, response intervals are assumed to reflect *uncertainty* around a respondent's best guess for the unknown value. For a latent interval truth, interval responses are assumed to represent *ambiguity* regarding the unknown interval (e.g., the consensus range of probabilities in the example of verbal quantifiers). In the example of judgments of risks, the *plausible range* of a particular risk might be of interest. The interval response again represents uncertainty, but now the uncertainty itself is the desired true quantity, i.e., an interval truth.

To facilitate the estimation of consensus intervals, we developed the Interval Truth Model (ITM), which combines and extends three previous contributions to the literature. First, the core of the model is the unidimensional consensus model by Anders et al. (2014), which uses a logit-normal distribution to model continuous bounded responses in $(0, 1)$. Second, we extend this model to two dimensions via a bivariate normal distribution, as previously implemented for unbounded responses by Mayer and Heck (2023). Third, we use the isometric log-ratio (ILR) transformation function (Smithson & Broomell, 2024) as an appropriate link function that connects the bivariate-normal model to the bounded interval responses.

We will explain the mathematical details of the ITM along with a Bayesian estimation method in Section 2 and consecutively present a simulation study for the computational evaluation of the model in Section 3. We will then apply the model in a reanalysis of the judgments of verbal quantifiers collected by Kloft and Heck (2024) in Section 4. Lastly, we will discuss implications, limitations, and directions for future research in Section 5.

2 Theory

2.1 The Interval Truth Model

In this section, we will introduce the notation for the data and the parameters. An overview of these declarations along with short explanations can be found in Appendix A. We assume that interval responses are measured on a response scale from 0 to 1 so that the lower and upper interval bounds are given as $0 \leq X^L \leq X^U \leq 1$. We first transform the data into a more generalizable compositional form, namely, a simplex with three components which sum to one:

$$\boldsymbol{X} = \left[X^L,\ X^U - X^L,\ 1 - X^U \right]^\top . \tag{1}$$

Since any of the three components in $\boldsymbol{X}$ can be zero, we need to add a padding constant c to the components to ensure that we can later apply a log-ratio transformation. After adding the constant, the compositional form is restored by dividing each element of the vector by the sum of all its elements. This procedure corresponds to the simple replacement strategy by Martín-Fernández et al. (2003):

$$\boldsymbol{Y} = \frac{1}{1+3c} \left(\boldsymbol{X} + c\,\mathbf{1} \right) \quad \text{with } c = 0.01, \tag{2}$$

where $\mathbf{1}$ is a vector of ones. We can skip this step if none of the components in any of the responses is zero.

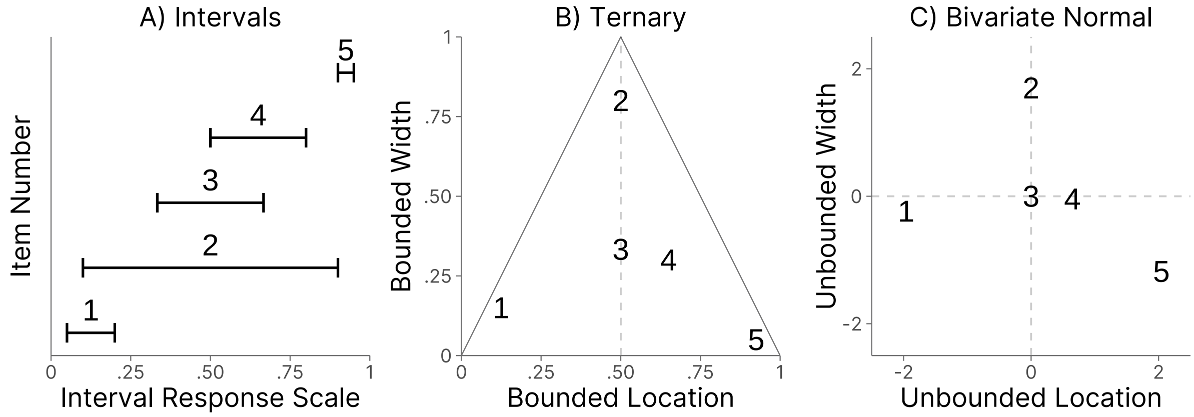
Next, we need to convert the intervals into a format better suited for our modeling framework, which assumes a bivariate normal distribution of response values. For this purpose, we apply a specific version of the isometric log-ratio (ILR) transformation function to $\boldsymbol{Y}$. This link function is tailored to the compositional form of interval responses (Smithson & Broomell, 2024):

$$\boldsymbol{Z} = \left[Z^{loc}, Z^{wid} \right] = \left[\sqrt{\frac{1}{2}} \log\left(\frac{Y_1}{Y_3} \right), \sqrt{\frac{2}{3}} \log\left(\frac{Y_2}{\sqrt{Y_1 Y_3}} \right) \right]^\top . \tag{3}$$

The transformation yields a vector $\boldsymbol{Z} \in \mathbb{R}^2$ with two elements, Z^{loc} and Z^{wid} that correspond to the unbounded interval location and interval width, respectively. The unbounded interval location Z^{loc} compares the size of the left component Y_1, defined by

Figure 2

Illustration of the Multivariate Logit Transformation



Note. The five observed response intervals are: Interval 1 $= [.05, .20]$, Interval 2 $= [.10, .90]$, Interval 3 $= \left[\frac{1}{3}, \frac{2}{3}\right]$, Interval 4 $= [.50, .80]$, Interval 5 $= [.90, .95]$.

the left response scale limit and the lower bound of the response interval, against the size of the right component Y_3, defined by the upper bound of the response interval. The unbounded interval width Z^{wid} compares the middle component Y_2, i.e., the observed interval width, to the geometric mean of the left and the right component, i.e., $\sqrt{Y_1 Y_3}$. The geometric mean in the denominator is used to scale the interval width relative to the interval location in the unbounded space. Therefore, a response interval of a particular width will be transformed into an unbounded interval of a greater width if the interval location is placed close to one of the response scale limits, compared to being placed near the center of the response scale. For example, the response interval $[.80, .90]$ with a width of .10 at the off-center location of .85 yields the transformed width $Z^{wid} = -0.85$. If an interval with the same observed width is placed near the center, e.g., $[.40, .50]$ with the location .45, it will yield the transformed width $Z^{wid} = -1.23$, which is considerably smaller.

Figure 2 illustrates the isometric log-ratio transformation for five response intervals. Interval 3 divides the response scale into a composition of three equally sized components (Panel A) and corresponds to the origin of the transformed, unbounded space in Panel C. For the unbounded location dimension (x-axis), the origin in the unbounded space in Panel C maps to the center of the bounded response scale in Panel B. Hence, unbounded locations of zero can be interpreted as neutral responses if the response scale is bipolar. In contrast, the origin on the unbounded width dimension (y-axis) does not have such a clear, substantive interpretation, since there is no such thing as a "neutral" width, at least not in the applications we are aware of.

Using the isometric log-ratio transformation as a link function, we can extend the model by Anders et al. (2014) to the two-dimensional case, similar to the model for geographical judgments by Mayer and Heck (2023). We choose this approach using the logit-link because it provides a more flexible alternative to the approach using a Dirichlet distribution (see Kloft et al., 2023, for an IRT model using this approach). Whereas the Dirichlet approach offers only one common variance parameter for both dimensions, the bivariate logit-normal distribution allows us to assume separate variance parameters for the unbounded location and width dimensions.

Next, we consider the bivariate, logit-transformed responses $\boldsymbol{Z}_{ij}$ of respondent $i = 1, \ldots, I$ (number of respondents) to item $j = 1, \ldots, J$ (number of items). We assume the following data-generating mechanism for a given interval response: Respondent i has a latent appraisal $[A_{ij}^{loc}, A_{ij}^{wid}]^\top \in \mathbb{R}^2$ for the item j based on the latent cultural consensus interval $[T_j^{loc}, T_j^{wid}]^\top \in \mathbb{R}^2$. This latent appraisal contains some error, depending on the proficiency of the person $[E_i^{loc}, E_i^{wid}]^\top \in \mathbb{R}_+^2$ and on the discernibility of the latent consensus for the particular item $[\lambda_j^{loc}, \lambda_j^{wid}]^\top \in \mathbb{R}_+^2$. Departing from previously developed CCT models (e.g., Anders et al., 2014), we inverted these parameters. Hence, higher values of proficiency and discernibility lead to higher precision of the latent appraisal, and thus, to observed response intervals that are closer to the latent true consensus interval. Moreover, we assume an item-specific correlation ω_j between the errors on the two dimensions (Mayer & Heck, 2023). Assuming a bivariate normal distribution of errors, the appraisal is centered on the latent cultural consensus with an added disturbance governed by person and item characteristics:

$$\begin{aligned} \begin{bmatrix} A_{ij}^{loc} \\ A_{ij}^{wid} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} T_j^{loc} \\ T_j^{wid} \end{bmatrix}, \boldsymbol{\Sigma}_{ij}^{A} \right) \quad \text{with} \quad \boldsymbol{\Sigma}_{ij}^{A} = \operatorname{diag}(\boldsymbol{\sigma}_{ij}^{A})\, \boldsymbol{\Omega}_j \operatorname{diag}(\boldsymbol{\sigma}_{ij}^{A}), \\ \boldsymbol{\sigma}_{ij}^{A} &= \begin{bmatrix} \frac{1}{E_i^{loc}\lambda_j^{loc}} \\ \frac{1}{E_i^{wid}\lambda_j^{wid}} \end{bmatrix}, \quad \boldsymbol{\Omega}_j = \begin{bmatrix} 1 & \omega_j \\ \omega_j & 1 \end{bmatrix}. \end{aligned} \tag{4}$$

The latent appraisal is further influenced by the respondent's scaling bias $a_i^{loc} \in \mathbb{R}_+^2$ and shifting biases $[b_i^{loc}, b_i^{wid}]^\top \in \mathbb{R}^2$:

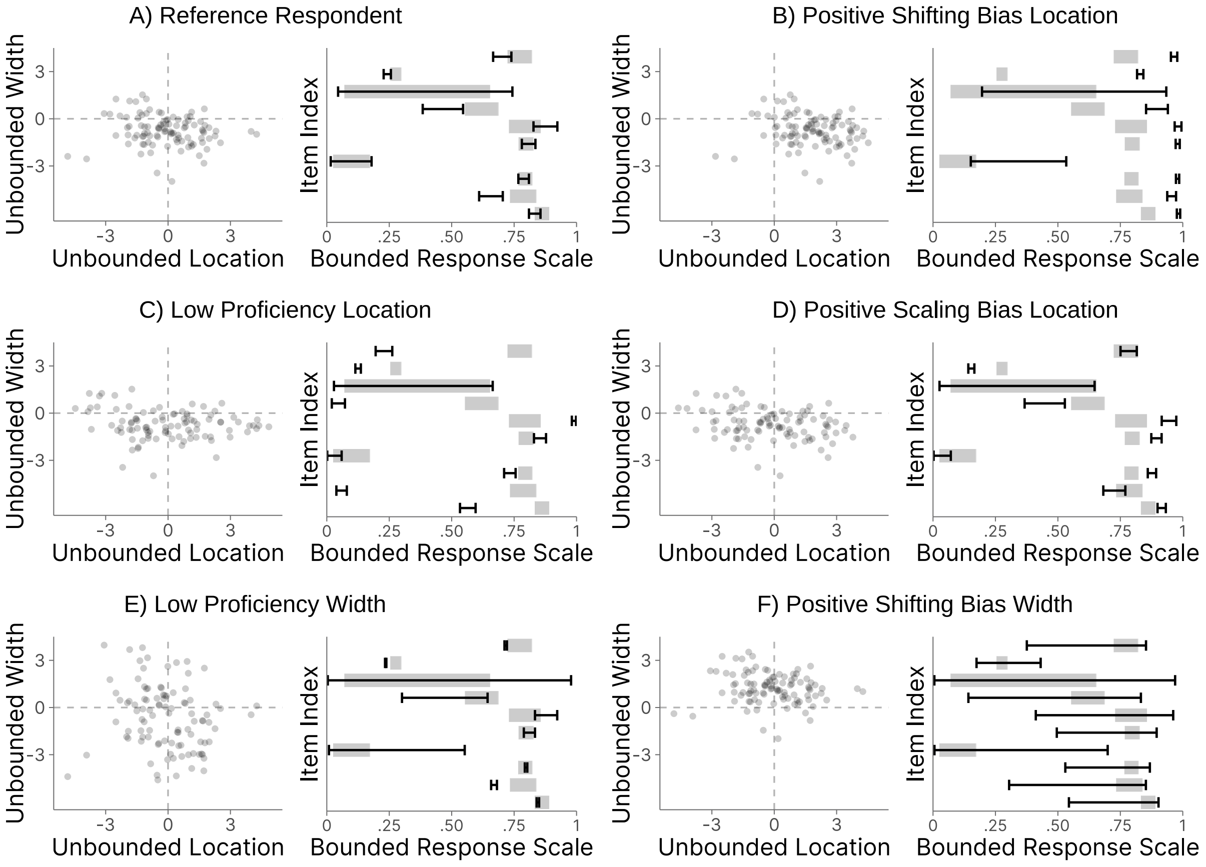
$$\boldsymbol{Z}_{ij} = \left[A_{ij}^{loc}\, a_i^{loc} + b_i^{loc},\ A_{ij}^{wid} + b_i^{wid}\right]^\top. \tag{5}$$

The two shifting biases are directional response biases and add a constant to each dimension of the latent appraisal, or, more technically, to the expected location and width. This corresponds to a respondent's tendency to systematically under- or overestimate all locations or all widths of the consensus intervals. The scaling bias corresponds to an extremity response bias, which pushes all observed responses of a person away from zero if $a_i^{loc} > 1$ or pulls them towards zero if $a_i^{loc} < 1$. As explained above, the origin $T_j^{wid} = 0$ of the width dimension is not a substantively meaningful anchor. We thus specify a scaling bias only for the location dimension, where respondents may scale all interval locations away from or towards the center of the bounded response scale. Since the appraisal of the interval location A_{ij}^{loc} consists of the consensus location plus an error, the scaling bias does not only influence the expected interval location but also the residual variance, i.e., the precision of the latent appraisal. In the full model, it is therefore necessary to ensure that the scaling bias parameter is included not only in the mean but also in the variance of the normal distribution:

$$\begin{aligned} \boldsymbol{Z}_{ij} &\sim \mathcal{N}(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}), \\ \boldsymbol{\mu}_{ij} &= \left[T_j^{loc} a_i^{loc} + b_i^{loc},\ T_j^{wid} + b_i^{wid}\right]^\top, \\ \boldsymbol{\Sigma}_{ij} &= \operatorname{diag}(\boldsymbol{\sigma}_{ij})\, \boldsymbol{\Omega}_j \operatorname{diag}(\boldsymbol{\sigma}_{ij}), \\ \boldsymbol{\sigma}_{ij} &= \begin{bmatrix} \frac{a_i^{loc}}{E_i^{loc}\lambda_j^{loc}} \\ \frac{1}{E_i^{wid}\lambda_j^{wid}} \end{bmatrix}, \quad \boldsymbol{\Omega}_j = \begin{bmatrix} 1 & \omega_j \\ \omega_j & 1 \end{bmatrix}. \end{aligned} \tag{6}$$

Figure 3

Illustration of the Person Parameters in the Interval Truth Model



Note. The scatterplots in the left-hand subpanels show the responses of one respondent to 100 randomly drawn items on the unbounded bivariate scale. The right-hand subpanels show the responses to ten selected items on the bounded interval-response scale (black intervals). The true consensus intervals, which are identical across all plots, are shown as gray, shaded bars in the background of the responses for the particular parameter setting.

The model can easily be modified by omitting bias parameters that are not relevant for certain applications (see also Section 4.1).

Figure 3 shows the isolated influence of each person parameter when all remaining parameters are held constant. Figure 3A contains the reference setup, namely, a respondent with high proficiency $[E_i^{loc}, E_i^{wid}]^\top = [3, 3]^\top$ and without any response biases, i.e., there is no scaling or shifting of the latent appraisal as $a_i^{loc} = 1$, $b_i^{loc} = 0$, and $b_i^{wid} = 0$.

While the true latent consensuses and latent appraisal errors were drawn randomly for the reference respondent in Figure 3A, each of the other panels shows the effect of manipulating one parameter and re-computing the responses accordingly. Hence, one can interpret the resulting changes in response patterns by comparing each panel to Figure 3A. Each panel consists of two subpanels that show different representations of the same response pattern. The subpanels on the left show 100 response intervals randomly drawn from the particular parameter setting on the unbounded two-dimensional scale. The subpanels on the right show ten exemplary response intervals for the same example items across all panels. Horizontal black segments represent the observed interval responses on the bounded scale for these items. Bold gray bars in the background show the respective latent consensus intervals, which serve as a reference for interpreting the response patterns. The latent consensus intervals are identical across all panels.

For a person with a low proficiency concerning interval locations, Figure 3C shows that response intervals move away from the true consensus interval. Similarly, for a person with a low proficiency concerning interval widths, Figure 3E shows that the widths of response intervals become less similar to the widths of the true consensus intervals. The effect of inducing a large scaling bias for interval locations (Figure 3D) is that the response intervals are shifted away from the center of the scale. With a large positive location shifting bias, shown in Figure 3B, all response intervals move to the right. Similarly, for a large positive shifting bias concerning interval widths, Figure 3F shows that all response intervals are greatly expanded in width.

The model needs to be fitted to data in order to estimate the latent consensus interval $[T_j^{loc}, T_j^{wid}]^\top$. For a meaningful interpretation of this estimate, we can convert the unbounded interval back to the original, bounded response scale. First, we convert the two-dimensional logit values to the compositional format via the inverse isometric log-ratio function, and second, we undo the smoothing initially applied by adding a padding constant:

$$\begin{aligned}\boldsymbol{T}_j^* &= (1+3c)\left[\frac{\exp\left(\sqrt{2}\,T_j^{loc}\right)}{\Sigma},\ \frac{\exp\left(\sqrt{\frac{3}{2}}\,T_j^{wid}+\frac{T_j^{loc}}{\sqrt{2}}\right)}{\Sigma},\ \frac{1}{\Sigma}\right]^\top - c\,\mathbf{1}, \\ \text{with}\quad \Sigma &= \exp\left(\sqrt{2}\,T_j^{loc}\right)+\exp\left(\sqrt{\frac{3}{2}}\,T_j^{wid}+\frac{T_j^{loc}}{\sqrt{2}}\right)+1,\end{aligned} \tag{7}$$

where, again, $c = .01$ and $\mathbf{1}$ is the vector of ones. Third, we compute the actual interval boundaries on the bounded scale from 0 to 1:

$$\left[T_j^L, T_j^U\right]^\top = \left[T_{j1}^*,\ T_{j1}^* + T_{j2}^*\right]. \tag{8}$$

The interval formed by $[T_j^L, T_j^U]^\top$ is the estimated consensus for the specific item, which we are ultimately interested in.

2.2 Bayesian Estimation

We estimate the model in a Bayesian hierarchical modeling framework (Kruschke & Vanpaemel, 2015). An illustration of the prior distributions can be found in the supplementary materials in the OSF repository. We are mainly interested in the latent consensus $[T_j^{loc}, T_j^{wid}]^\top$. First, since we know that very wide interval widths are highly unlikely and also not meaningful in most use cases, we assign a weakly informative prior on the true interval widths on the bounded scale:

$$T_j^{wid(0,1)} \sim \text{Beta}(1.2, 3) \tag{9}$$

This prior has an expected value of .29 and a mode of .09 and therefore reflects our beliefs about the marginal true interval widths more adequately than a uniform prior.

Second, conditional on a particular interval width, we do not believe that particular interval locations are more likely than others. Therefore, we assign an uninformative prior to an auxiliary shifting weight parameter s_j, which is subsequently used to compute the actual true locations on the bounded scale:

$$\begin{aligned} s_j &\sim \text{Beta}(1, 1), \\ T_j^{loc(0,1)} &= s_j(1 - T_j^{wid(0,1)}) + \frac{T_j^{wid(0,1)}}{2}. \end{aligned} \tag{10}$$

This means that for a given interval width, we take what is left of the response scale and multiply it by the shifting weight s_j, which results in the lower bound for this particular interval. To arrive at the interval location, we add half of the respective interval width to this lower bound.

Third, we transform the true interval from the bounded simplex to the unbounded bivariate scale via the isometric log-ratio function:

$$\boldsymbol{T}^* = \text{ILR}\left(\left[T_j^{loc(0,1)} - \frac{T_j^{wid(0,1)}}{2},\ T_j^{wid(0,1)},\ T_j^{loc(0,1)} + \frac{T_j^{wid(0,1)}}{2}\right]^\top\right). \tag{11}$$

Similarly, we could have applied an uninformative prior directly on the simplex via a Dirichlet distribution. An implementation of this prior can also be found in the OSF repository:

$$\text{ILR}^{-1}(\boldsymbol{T}^*) \sim \text{Dirichlet}(1, 1, 1). \tag{12}$$

Assigning beta priors to the parameters on the bounded scale instead of a bivariate normal prior on the unbounded scale (a) allows us to incorporate our domain knowledge more intuitively, and (b) additionally helps with the stability of the model estimation.

The person proficiency parameters $[E_i^{loc}, E_i^{wid}]^\top$ have weakly informative priors on both the means and the variances (Table 1, column 1). The priors are on the log-scale to ensure positive values. We are also interested in the relationship between a respondent's proficiency in the location dimension and their proficiency in the width dimension, and therefore assign a bivariate normal prior. Similarly, we assign the same priors to the item discernibilities $[\lambda_j^{loc}, \lambda_j^{wid}]^\top$ (Table 1, column 2). The only difference is that we fix the mean vector μ_λ to zero to render the person proficiency parameters identifiable (see also Anders et al., 2014).

For the remaining person parameters, namely, the scaling and the shifting biases, we also assign weakly informative priors. In doing so, we impose certain restrictions on

Table 1

Default Prior Distributions for the Interval Truth Model

Person proficiency $\mathbf{E}_i$	Item difficulty λ_j
$[\log(E_i^{loc}), \log(E_i^{wid})]^\top \sim \mathcal{N}(\boldsymbol{\mu}_E, \boldsymbol{\Sigma}_E)$	$[\log(\lambda_j^{loc}), \log(\lambda_j^{wid})]^\top \sim \mathcal{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)$
$\boldsymbol{\mu}_E \sim \mathcal{N}(0, 1)$	$\boldsymbol{\mu}_\lambda = \mathbf{0}$
$\boldsymbol{\Sigma}_E = \mathrm{diag}(\boldsymbol{\sigma}_E)\, \boldsymbol{\Omega}_E\, \mathrm{diag}(\boldsymbol{\sigma}_E)$	$\boldsymbol{\Sigma}_\lambda = \mathrm{diag}(\boldsymbol{\sigma}_\lambda)\, \boldsymbol{\Omega}_\lambda\, \mathrm{diag}(\boldsymbol{\sigma}_\lambda)$
$\boldsymbol{\Omega}_E = \boldsymbol{\Omega}_{EL}\boldsymbol{\Omega}_{EL}^T$	$\boldsymbol{\Omega}_\lambda = \boldsymbol{\Omega}_{\lambda L}\boldsymbol{\Omega}_{\lambda L}^T,$
$\boldsymbol{\Omega}_{EL} \sim \text{LKJ-Cholesky}(2)$	$\boldsymbol{\Omega}_{\lambda L} \sim \text{LKJ-Cholesky}(2)$
$\log(\boldsymbol{\sigma}_E) \sim \mathcal{N}(\log[0.5], 0.5)$	$\log(\boldsymbol{\sigma}_\lambda) \sim \mathcal{N}(\log[0.5], 0.5)$

the means for reasons of identifiability (see also Anders et al., 2014):

$$\begin{aligned}
\log(a_i^{loc}) &\sim \mathcal{N}(0, \sigma_{a^{loc}}), \\
b_i^{loc} &\sim \mathcal{N}(0, \sigma_{b^{loc}}), \\
b_i^{wid} &\sim \mathcal{N}(0, \sigma_{b^{wid}}), \\
\log(\sigma_{a^{loc}}) &\sim \mathcal{N}(\log[0.5], 0.5), \\
\log(\sigma_{b^{loc}}), \log(\sigma_{b^{wid}}) &\sim \mathcal{N}(\log[0.5], 1).
\end{aligned} \tag{13}$$

The mean of the shifting bias parameters $[b_i^{loc}, b_i^{wid}]^\top$ is fixed to $\mathbf{0}$ to make the model identifiable regarding the estimated mean of the latent true locations and widths. Analogously, the mean of the scaling bias parameters a_i^{loc} is fixed to 1 to render the model identifiable regarding the estimated mean of the proficiency parameters $[E_i^{loc}, E_i^{wid}]^\top$.

Finally, we assign weakly informative priors to the residual correlation between interval location and width via a scaled beta distribution:

$$\frac{\omega_j + 1}{2} \sim \mathrm{Beta}(2, 2). \tag{14}$$

3 Simulation Study

The simulation study was preregistered using the ADEMP preregistration template by Siepe et al. (2023) to specify the Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures. The pre-registration is available at the Open Science Framework (https://osf.io/nd5wg). After running the simulation with the pre-registered settings, we found that some conditions had a lot of problematic model fits. We therefore decided to re-work the parameterization and priors of the model for more stable model estimation, and subsequently re-ran the simulation. We will indicate the changes to the preregistered settings where applicable. We further provide all results of the original simulation study in the supplementary materials in the OSF repository. The main results did not change, as both the best-performing model per condition and the overall trends of the performance measures remained the same. The simulation study was carried out in the programming environment `R` Version 4.4.1 (R Core Team, 2023) on a Linux machine with an Ubuntu 22.04.4 LTS distribution. We provide a Dockerfile to reproduce our main results. We used the following R packages 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` packages (Gabry & Mahr, 2024) for handling and visualizing MCMC output. Additional packages used for data wrangling and minor tasks are provided in the supplementary materials in the OSF repository.

3.1 Aims

The simulation study aimed to explore the estimation performance of the interval truth model (ITM) concerning bias and mean-squared error of parameter estimates in realistic scenarios of use. The main target estimates were the latent true interval location and width $[T_j^{loc}, T_j^{wid}]^\top$. We also tracked the performance of the other parameters, except for the hyperparameters. We compared the model estimates of the latent consensus intervals for each item against simple means of the logit-transformed responses as a simple competitor model (i.e., wisdom of crowds; Surowiecki, 2004). Given that the data are generated from our model, we expected the model estimates to perform better than simple means. If that was not the case, the added complexity of our model may not be worth the effort compared to relying on simpler descriptive aggregation strategies. We further expected that larger numbers of respondents would lead to better performance of item parameters, and, vice versa, that larger numbers of items would lead to better performance of person parameters.

In addition to the main simulation study, we conducted a preliminary simulation study to test the isometric log-ratio function against an alternative log-ratio transformation, which is based on a stick-breaking procedure (see Smithson & Broomell, 2024). We were interested in checking the robustness of the two link functions regarding model misspecification. We generated data with one fixed combination of 200 respondents and 30 items and only varied the link function used to simulate the data, resulting in two conditions. Each model was fitted to the data using the data-generating link function as well as the non-data-generating link function. We report the full results of this preliminary simulation study in the supplementary materials in the OSF repository.

3.2 Data Generation

We randomly generated data from the model described in Section 2.1. We varied the following factors in a fully factorial manner:

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

This yielded 16 conditions. The numbers of respondents and items were selected to cover a range of practically relevant applications. There may be scenarios with only a few items and few expert raters, for instance, when a company has ten expert employees who judge the risk of a security breach for five software components. In other scenarios, large numbers of raters and items might be available, for instance, in a forecasting challenge.

In all conditions, the true, data-generating parameters were randomly drawn for each repetition. We used the model described in Section 2.1 as the data-generating mechanism for each interval response $\boldsymbol{Z}_{ij}$ of respondent i to item j on the unbounded scale. To obtain manifest interval responses in the bounded simplex space, we first

Table 2

Distributions of Hyperparameters Used for Data Generation

Interpretation	Parameter	Distribution
Items		
True location	T_j^{loc}	$\mathcal{N}(0, 0.81)$
True width	T_j^{wid}	$\mathcal{N}(-0.57, 0.65)$
Location discernibility	$-\log(\lambda_j^{loc})$	$\mathcal{N}(0, 0.3)$
Width discernibility	$-\log(\lambda_j^{wid})$	$\mathcal{N}(0, 0.3)$
Residual correlation	ω_j	0
Respondents		
Location proficiency	$-\log(E_i^{loc})$	$\mathcal{N}(\log[0.81], 0.3)$
Width proficiency	$-\log(E_i^{wid})$	$\mathcal{N}(\log[0.65], 0.3)$
Location scaling bias	$\log(a_i^{loc})$	$\mathcal{N}(0, 0.3)$
Location shifting bias	b_i^{loc}	$\mathcal{N}(0, 0.27)$
Width shifting bias	b_i^{wid}	$\mathcal{N}(0, 0.22)$

Note. For the parameters E_j and λ_j we defined the distributions on the inverted scale, i.e., on the variance instead of the precision scale.

transformed the unbounded interval response $\boldsymbol{Z}_{ij}$ using the inverse isometric log-ratio function (Smithson & Broomell, 2024). In the model estimation step, the data were then transformed back to the unbounded space using the isometric log-ratio function (Equation 7 with $c = 0$, see also Smithson & Broomell, 2024). This back-and-forth transformation is a redundant step for our main simulation study, where the same transformation was used for data generation and model estimation. However, for our preliminary simulation study investigating the performance of different link functions, this is a crucial step required to cross-fit a model version with one link function to the data generated with the respective other link function.

Table 2 lists all hyperparameter values used for generating person- and item-specific model parameters. The preregistration protocol contains a detailed justification of these values. Overall, we aimed to generate plausible distributions of manifest response intervals. We derived the hyperparameters from actual response intervals representing typical or extreme responses. For the true mean of location and width, we used the values resulting from the logit-transformed interval $[.40, .60]$. For the standard deviation of the true location, we used the interval $[.98, .99]$ transformed to the bivariate space. We declared the resulting unbounded value as the point that is four standard deviations away from the unbounded mean location, i.e., an extreme value in the unbounded space. We further calculated the standard deviation for the unbounded true location by dividing the difference between this extreme value and the mean of the true location by four. Analogously, for the standard deviation of the true width, we used the interval $[.495, .505]$). We simulated the true location and width parameters from normal distributions since all parameters were defined on the latent, unbounded scale.

Due to the large computational demand of our simulation study, we decided on the number of repetitions as follows: We aimed for a Monte Carlo standard error (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 500 (pre-registration: $1,000$) repetitions computationally reasonable. After 500 repetitions, we checked the MCSEs in all conditions. If they had not met the above criterion, we would have incrementally added repetitions in steps of 250 until they did. The MCSEs in all conditions had met the criterion after 500 repetitions, with the largest MCSE of the absolute bias being 0.005 in one condition.

Further details can be found in the preregistration and in the supplementary materials in the OSF repository, where we illustrate the distributions of parameters and responses as well as the recovery of one set of data-generating parameters.

3.3 Method

We estimated 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 `rstan` (Stan Development Team, 2024). For the Bayesian estimation, we used the priors described in

Section 2.2. The only exception was that we did not use two-dimensional, multivariate priors for $[E_i^{loc}, E_i^{wid}]^\top$ and $[\lambda_j^{loc}, \lambda_j^{wid}]^\top$, but rather independent univariate prior distributions. For each repetition, we ran four chains of Stan's Hamiltonian Monte Carlo sampler (Betancourt, 2018) with 500 warm-up samples not used for analyses and 500 (preregistration: $1,000$) samples for the computation of parameter estimates, which yielded $2,000$ samples per parameter. Given the results for the convergence diagnostics shown below, we deemed this number sufficient. The `adapt_delta` parameter was set to 0.999 for conditions with a number of total simulated responses $\leq 2,000$, and to 0.9 for the conditions with a greater number (preregistration: 0.9 for all conditions). We changed this setting because in our earlier simulations we had encountered issues with divergent transitions in conditions with low numbers of responses. The range of the initial values of the sampling algorithm for the unbounded parameters was set to $[-0.1, 0.1]$.

3.4 Performance Measures

Our primary performance measure was the absolute bias of the latent, unbounded consensus interval location and width $[T_j^{loc}, T_j^{wid}]^\top$, which we defined as follows:

$$\widehat{\text{AbsBias}} = \frac{\sum_{n=1}^{N}\sum_{j=1}^{J} 0.5\left(\left|\hat{T}_{nj}^{loc} - T_{nj}^{loc}\right| + \left|\hat{T}_{nj}^{wid} - T_{nj}^{wid}\right|\right)}{N \times J},$$

where J is the number of items in a specific condition and N is the number of repetitions of the simulation. We computed the mean of the (absolute) bias of location and width jointly because we expected that there could be a compensatory effect concerning the accuracy of estimates. We also computed the absolute bias for both dimensions separately for illustration purposes below. We additionally calculated the location- and width-specific biases and visualized them jointly in a scatter plot to assess such potential compensatory effects (see the supplementary materials in the OSF repository).

We additionally calculated the mean squared error (MSE) for the bivariate vector $[T_j^{loc}, T_j^{wid}]^\top$ of the latent, unbounded consensus intervals:

$$\widehat{\text{MSE}} = \frac{\sum_{n=1}^{N}\sum_{j=1}^{J} 0.5\left(\left(\hat{T}_{nj}^{loc} - T_{nj}^{loc}\right)^2 + \left(\hat{T}_{nj}^{wid} - T_{nj}^{wid}\right)^2\right)}{N \times J}.$$

We also calculated the MSE for the location and width individually. We estimated the MCSE of these performance measures via bootstrapping. We further tallied the number of non-converged simulation repetitions.

3.5 Results

3.5.1 Preliminary Study: Link Functions

Our preliminary simulation that compared two alternative link functions showed the superiority of the isometric log-ratio transformation over the stick-breaking transformation. Especially in the case of cross-fitting the model to the data generated by the respective other link function, the isometric log-ratio transformation was more robust to this specific type of model misspecification.

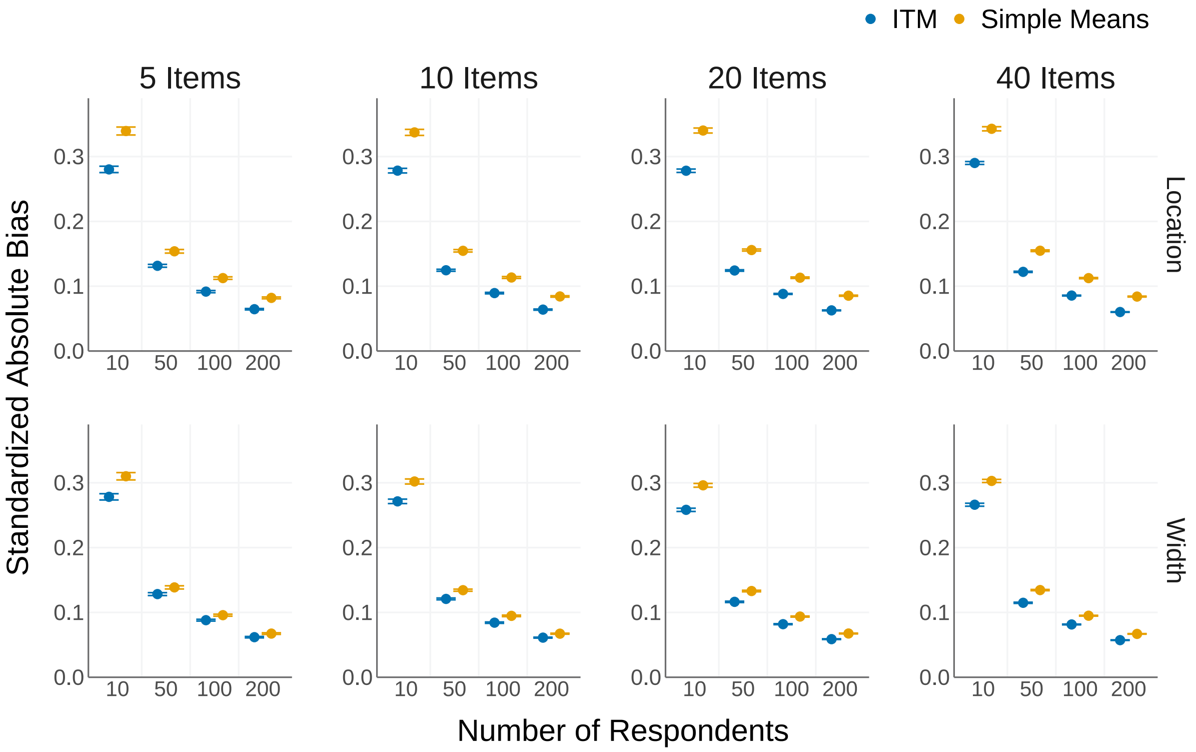
3.5.2 Main Study: Recovery of Latent Consensus Intervals

All repetitions of the simulation study finished without error. The average $\hat{R}$ across all repetitions and conditions was 1.002. In 13 of 16 simulation conditions, we observed no divergent transitions. In the "worst" condition with ten respondents and five items, 1.2% of all models contained at least one divergent transition. The models with divergent transitions in this condition contained on average 26.7 divergent transitions. Overall, these results imply good convergence in almost all repetitions. This indicates that the model can even be estimated in edge cases with a low number of items and respondents, where the performance benefit compared to the aggregation via simple means is particularly large. Additional results on convergence metrics are available in the supplementary materials in the OSF repository.

We visualized the absolute bias of the latent interval location and width in Figure 4. The true locations had a higher standard deviation (0.81) compared to the true widths (0.65). Therefore, we divided the absolute bias by the true standard deviations of the respective parameters for ease of interpretation in the figure. The unstandardized performance measures are available in the OSF repository. In all simulation conditions, the ITM has a lower absolute bias averaged over location and

Figure 4

Absolute Bias of Interval Location and Width.



Note. This figure shows the standardized absolute bias (y-axis) of the interval location (upper row) and width (lower row) for different numbers of items (columns) and numbers of respondents (x-axis). The standardized absolute bias was obtained by dividing the condition-wise absolute bias by the true standard deviation of the location or width. Error bars indicate ±1 MCSE. Some MCSEs are so small that the upper and lower error bars are indiscernible.

width parameters than the simple means. As expected, there is a notable effect of the number of respondents, with a considerably lower bias for higher sample sizes. Increasing the number of respondents from 10 to 50 roughly corresponds to halving the absolute bias for all conditions. The size of the performance difference between the ITM and the simple means remains fairly similar for sample sizes 50 to 200. A larger number of items slightly improves the performance of the ITM regarding the recovery of true intervals, but this effect is weaker than the effect of the number of respondents. The standardized absolute bias is very similar for the location and width dimensions, which means that both dimensions can be estimated similarly well. We chose to plot both dimensions separately here to illustrate this point. The combined absolute bias, which we defined above, shows a virtually identical pattern of results.

In the supplementary materials in the OSF repository, we present additional simulation results. These show that the mean squared error follows a qualitatively very similar pattern to the results of the absolute bias. For all conditions, the ITM had a better performance concerning the MSE than the simple means. Further, in simulation repetitions with a higher bias of the location, the bias of the width tended to be higher as well. Thus, we did not observe evidence for compensatory behavior, where an accurate estimation of one dimension would be associated with a poorer estimation of the respective other dimension.

The results of our simulation study indicate that the interval truth model performed better than simple means in all conditions we studied. The difference between both approaches became smaller with a larger number of respondents. The number of items did not have a strong influence on the results regarding the true intervals. This is not surprising because our performance measures are aggregated across the item parameters. However, the small increases in performance may be mediated by the increased accuracy of person parameters in conditions with larger numbers of items. The more accurate person parameters will in turn lead to more accurate estimates of the latent consensus intervals. As we standardized the absolute bias, the results can be interpreted as fractions of the true standard deviation, indicating a satisfactory performance of the ITM.

4 Empirical Example: Verbal Quantifiers

To demonstrate the application of the interval truth model (ITM), we reanalyze data by Kloft and Heck (2024). Participants provided judgments for verbal quantifiers such as “seldom” or “often” using the dual range slider response format (see Figure 1). For each verbal quantifier, respondents had to assign an interval of probabilities ranging from 0% to 100% according to the probability that an event described in this way would occur. The full analysis is available in the supplementary materials in the OSF repository.

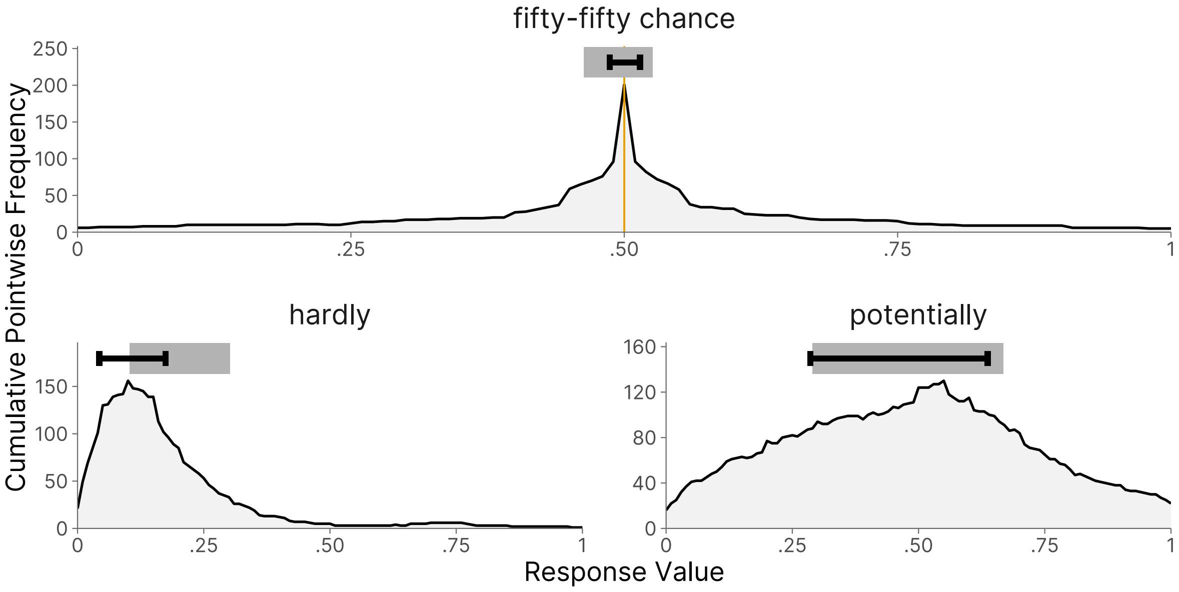
4.1 Model Modification and Estimation

Not all parameters described in Section 2 yielded useful estimates in an initial fit of the full model. Specifically, the estimates for the respondents’ response bias parameters b_i^{loc} (systematic shifts in the location dimension) did not differ meaningfully as indicated by a variance close to zero. We therefore simplified the model by excluding these parameters.

We estimated the model using the same software as in the simulation study but on a Windows machine. Information on the computational environment is provided in the supplementary materials in the OSF repository. For Bayesian inference, we used the priors described in Section 2.2. We ran four chains of Stan’s Hamiltonian Monte Carlo sampler (Betancourt, 2018) with 500 warm-up samples not used for analyses and $1,000$ samples for the computation of parameter estimates, which yielded $4,000$ samples per parameter. The `adapt_delta` parameter was set to 0.8 and the range of the initial values of the sampling algorithm for the unbounded parameters was set to $[-0.1, 0.1]$. Convergence was assessed via the $\hat{R}$ statistic (Vehtari et al., 2021), which was below 1.02 for all parameters. We provide posterior predictive checks in the online supplementary materials.

Figure 5

Estimated Consensus Intervals for Verbal Quantifiers



Note. Black horizontal interval: Consensus interval estimated by the interval truth model. Gray horizontal bar: Typical interval computed based on the mean location and mean width of the observed, logit-transformed response intervals.

4.2 Model Results

Figure 5 presents four examples of estimated intervals (black horizontal intervals) that each resemble the cultural consensus of the sampled respondents, jointly with a simple mean of logit-transformed interval responses (gray horizontal bars) and pointwise cumulative frequencies of the empirical interval responses (black density lines). The "fifty-fifty chance" item (top left) was one of the control items in the study, for which respondents were expected to answer with narrow intervals placed in the center of the response scale. The estimated consensus interval is centered on the correct reference value of 50% and very narrow, reflecting the high precision of the verbal statement "fifty-fifty chance." A substantial proportion of response intervals are wider (as indicated by the density), but the consensus is still that "50:50" is a probability very close to 50%. In contrast, the simple mean interval has a slight bias towards the left side of the response scale and is also wider. For the item “hardly” (bottom left), the estimated consensus interval is more representative of the empirical distribution, while the simple mean interval is shifted towards the center of the scale. This suggests that the estimates are less influenced by the inwardly skewed outliers of the empirical responses than the simple means. The model estimates seem to be less influenced by extreme responses. The fourth item, “potentially”, gives an example of a case where the model estimate and the simple mean yielded a very similar aggregated interval.

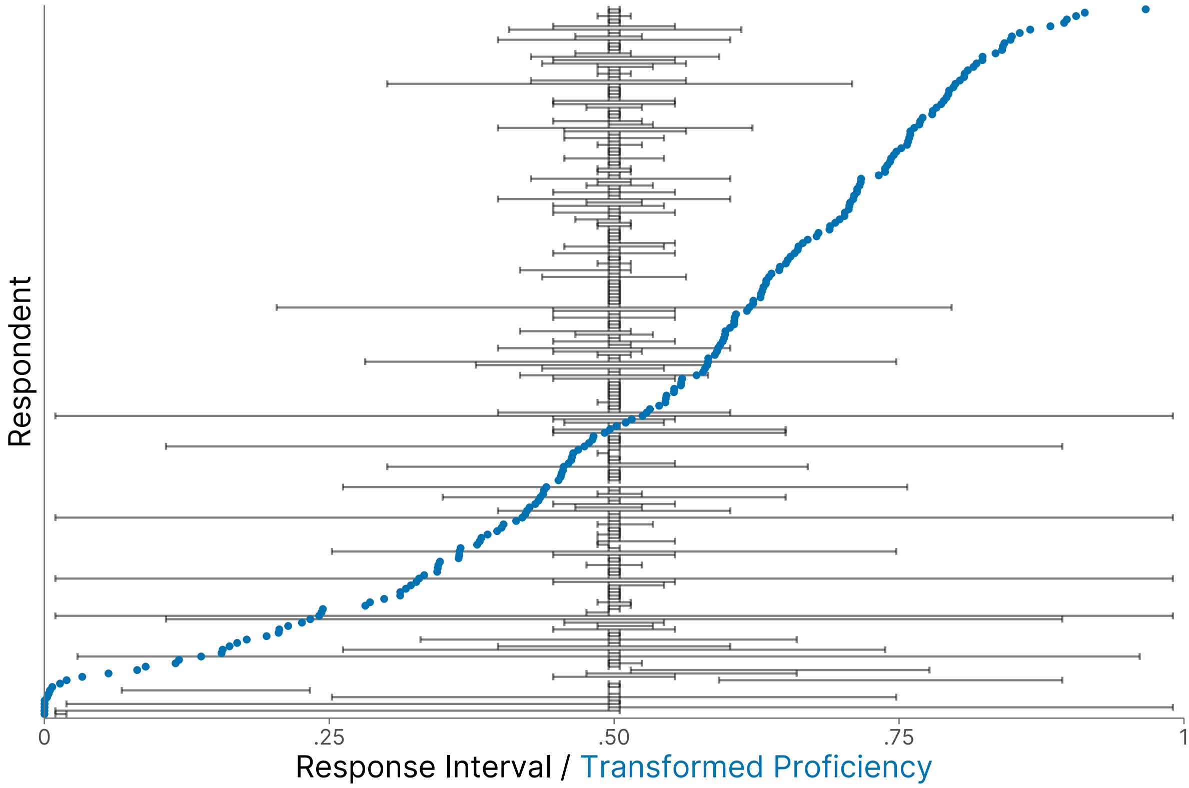
Figure 6 provides insight into how the estimated proficiency relates to empirical interval responses by overlaying empirical responses to the verbal quantifier “fifty-fifty chance” with proficiency estimates that are averaged over the location and width dimension. Respondents are ordered by their proficiency from high (top) to low (bottom). The respondents in the upper half of the y-axis mostly provided relatively narrow intervals located in the center of the response scale. In the lower half of the y-axis, several respondents provided very wide response intervals, which were also located in the center of the scale, for some due to their width. The respondents with the lowest proficiency at the bottom of the y-axis mostly failed to place the interval in the center of the response scale. Consequently, the proficiency estimates can be useful in diagnosing non-effortful responding. The model also enables us to down-weight the responses of unreliable respondents without having to exclude them from the data based on arbitrary filtering criteria.

The estimated correlation for the respondents’ proficiency between the location dimension and the width dimension (see also Table 2, Column 1) was $\hat{\rho}_E = .63$ (95% HDI $[.51, .74]$). Substantively, this means that respondents who answered highly consistent with respect to the interval location, i.e., the mean probability tied to a specific verbal quantifier, were also highly consistent regarding the interval widths, i.e., the variability in how the respective quantifier might be used.

For the items, discernibility of the true location and width (see also Table 2, Column 2) was correlated negatively with $\hat{\rho}_\lambda = -.47$ (95% HDI $[-.78, -.08]$). This correlation should be considered with caution since it is driven by the control items

Figure 6

Estimated Proficiencies vs. Empirical Interval Responses for "fifty-fifty chance"



Note. Black vertical bars: Empirical response intervals. Blue dots: Mean of standardized estimated location proficiency and width proficiency (posterior median) transformed to normal quantiles.

("never", "always", "fifty-fifty chance"). These had especially high location discernibility estimates above the mean and especially low width discernibility estimates. At the same time, all other items' location discernibility estimates were below the mean, and their width discernibility estimates were above the mean. We had selected these verbal quantifiers as controls because they have a clear implication for the typical probability that should be assigned to them, i.e., "never" = 0%, "always" = 100%, and "fifty-fifty chance" = 50%. The high location discernibility indicates that respondents overall interpreted these quantifiers in the assumed way. To check if the correlation of the discernibility dimensions was just due to the control items' influence, we re-fitted the model, excluding the three control items. As we expected, the correlation was no longer negative and even changed to a large positive value with $\hat{\rho}_\lambda = .81$ (95% HDI $[.48, .98]$). This means that items with an easy-to-detect location also tended to have a width that was easier to detect. At the same time, the correlation for the respondents' proficiency was reduced to $\hat{\rho}_E = .50$ (95% HDI $[.34, .65]$). In conclusion, we can use item parameters like discernibility to facilitate manipulation checks or to weed out poorly performing items.

5 Discussion

We proposed the Interval Truth Model (ITM), a cultural consensus model that can be applied to continuous, bounded interval responses from which consensus intervals are estimated. In a simulation study, we demonstrated that the consensus model performs better than simple averaged intervals. We also showed that the model can be estimated with as little data as five items and ten respondents. We further illustrated the application of the proposed model to empirical data, namely interval judgments of verbal quantifiers, and showed that the model can be used to detect and down-weight unreliable respondents.

The results of our simulation study showed that our choice of the isometric log-ratio transformation over the stick-breaking transformation was justified, as the former is more robust to model misspecification. The ITM showed good convergence and a better performance in terms of absolute bias and MSE than aggregation via simple means. Even in the "hardest" conditions with small numbers of individuals and items, divergent transitions occurred only in a small proportion of models. If such divergences should occur in empirical research, more informative priors should be applied. This is also a strength of the Bayesian estimation approach, which allows for incorporating prior knowledge about the consensus intervals and robust estimation, even in small datasets.

In the empirical example, the consensus intervals were centered on the true value in the case of "fifty-fifty chance" or on the mode of the distribution in the case of "hardly." Compared to the simple means, the consensus appeared to be more robust against extreme responses from individuals with low proficiency. It can therefore be used to obtain higher-quality estimates of a latent truth. While simple trim-and-average heuristics (Gaba et al., 2017; Lyon et al., 2015; Park & Budescu, 2015) could be useful in this regard, our model-based approach has the additional advantage of providing estimates for the proficiency of respondents and the discernibility of items. These estimates may be used for diagnostic purposes, as illustrated by the analysis of control items in the empirical example. Further, the ITM could be extended to an explanatory model, for example, by incorporating latent regressions.

While we confined ourselves to the simplest possible version of a consensus model for interval responses, there are several possibilities for extensions of this model, which we did not cover in the present article. The model is limited to only one latent true interval per item. There might be more than one latent truth across different groups, in other words, latent classes of respondents (see Anders et al., 2014). Also, we did not cover the case where the latent truth is a point, but the responses are collected with an interval format, as in forecasting (Gaba et al., 2017; Peeters & Wolk, 2017). The ITM might be used in such cases to derive consensus intervals for point truths to judge performance based on coverage of the true value. However, it is desirable to develop a specific model, which assumes that the latent consensus is a point, but estimates this point truth using interval responses. In such a case, the interval responses would reflect the uncertainty of respondents about a specific, unknown probability that a certain event will occur in the future. The development of these extended models was beyond the scope of this article and might be a promising avenue for future research.

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

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181. https://doi.org/10.1007/s11336-013-9382-9

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*(5), 316–332. https://doi.org/10.1016/j.jmp.2012.06.002

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*(1), 71–92. https://doi.org/10.1007/BF02294195

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*. https://doi.org/10.48550/arXiv.1701.02434

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for english words (anew): Instruction manual and affective ratings* (tech. rep.). Technical report C-1, The Center for Research in Psychophysiology, University of Florida.

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

Ellerby, Z., McCulloch, J., Wilson, M., & Wagner, C. (2020). Exploring how component factors and their uncertainty affect judgements of risk in cyber-security. In S. Nadjm-Tehrani (Ed.), *Critical information infrastructures security* (pp. 31–42). Springer International Publishing. https://doi.org/10.1007/978-3-030-37670-3_3

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

Gaba, A., Tsetlin, I., & Winkler, R. L. (2017). Combining interval forecasts. *Decision Analysis*, *14*(1), 1–20. https://doi.org/10.1287/deca.2016.0340

Gabry, J., Češnovar, R., & Johnson, A. (2023). *Cmdstanr: R interface to 'cmdstan'* (Version 0.8.1) [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/

Gersen, L. (2024). Leongersen/noUiSlider [Software]. https://github.com/leongersen/noUiSlider

Harris, A. J. L., Por, H.-H., & Broomell, S. B. (2017). Anchoring climate change communications. *Climatic Change*, *140*(3), 387–398. https://doi.org/10.1007/s10584-016-1859-y

Karelitz, T. M., & Budescu, D. V. (2004). You say "Probable" and I say "Likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied*, *10*(1), 25–41. https://doi.org/10.1037/1076-898X.10.1.25

Kloft, M., Hartmann, R., Voss, A., & Heck, D. W. (2023). The Dirichlet dual response model: An item response model for continuous bounded interval responses. *Psychometrika*. https://doi.org/10.1007/s11336-023-09924-7

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. *Educational and Psychological Measurement*. https://doi.org/https://doi.org/10.1177/00131644241283400

Kloft, M., Snijder, J.-P., & Heck, D. W. (2024). Measuring the variability of personality traits with interval responses: Psychometric properties of the dual-range slider response format. *Behavior Research Methods*. https://doi.org/10.3758/s13428-024-02394-4

Kruschke, J. K., & Vanpaemel, W. (2015, December). *Bayesian estimation in hierarchical models* (J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.13

Lyon, A., Wintle, B. C., & Burgman, M. (2015). Collective wisdom: Methods of confidence interval aggregation. *Journal of Business Research*, *68*(8), 1759–1767. https://doi.org/10.1016/j.jbusres.2014.08.012

Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, *35*(3), 253–278. https://doi.org/10.1023/A:1023866030544

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

Navarro, J., Wagner, C., Aickelin, U., Green, L., & Ashford, R. (2016). Exploring differences in interpretation of words essential in medical expert-patient communication. *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2157–2164. https://doi.org/10.1109/FUZZ-IEEE.2016.7737959

Park, S., & Budescu, D. V. (2015). Aggregating multiple probability intervals to improve calibration. *Judgment and Decision Making*, *10*(2), 14.

Peeters, R., & Wolk, L. (2017). Eliciting interval beliefs: An experimental study. *Plos one*, *12*(4), e0175163. https://doi.org/10.1371/journal.pone.0175163

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

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338.

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

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

Stan Development Team. (2024, October 1). *Rstan: R Interface to Stan*. Retrieved January 24, 2024, from https://mc-stan.org/cmdstanr/

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday & Co.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Appendix

Abbreviations and Parameter Interpretations

- CCT: Cultural Consensus Theory
- ITM: Interval Truth Model
- VAS: Visual Analog Scale
- DRS: Dual Range Slider
- MCMC / HMC: Markov Chain Monte Carlo / Hamiltonian Monte Carlo
- NUTS: No-U-Turn Sampler
- HDI: Highest Density Interval (for a given posterior distribution; Bayesian)
- $\widehat{R}$: Statistic for the diagnosis of MCMC convergence
- Data Declarations:
  - X^L, X^U: Lower and upper bound of interval response
  - $\mathbf{X}$: Interval response in its simplex representation
  - $\mathbf{Y}$: Interval response in its simplex representation after adding a padding constant
  - $\mathbf{Z} = [Z^{loc}, Z^{wid}]^\top$: Logit-transformed interval response
- Model Parameters of Interval Truth Model:
  - A_{ij}^{loc}, A_{ij}^{wid}: Respondent's latent appraisal of interval location and width
  - a_i^{loc}: Person scaling bias for the interval location
  - b_i^{loc}, b_i^{wid}: Person shifting bias for the interval location and width
  - E_i^{loc}, E_i^{wid}: Person proficiency to detect the true interval location and width
  - T_j^{loc}, T_j^{wid}: Latent true interval location and width
  - T_j^L, T_j^U: Latent true interval lower and upper boundary on the bounded response scale

- λ_j^{loc}, λ_j^{wid}: Item discernibility for the interval location and width
- ω_j: Residual correlation between location and width dimension