UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Schnuerch, M., Haaf, J. M., Sarafoglou, A., & Rouder, J. N. (2022). Meaningful Comparisons With Ordinal-Scale Items. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.38594

Methodology and Research Practice

Meaningful Comparisons With Ordinal-Scale Items

Martin Schnuerch[1] [a], Julia M. Haaf[2], Alexandra Sarafoglou[2], Jeffrey N. Rouder[3]

[1] Department of Psychology, University of Mannheim, Mannheim, Germany, [2] University of Amsterdam, Amsterdam, Netherlands, [3] University of California, Irvine, CA, US; University of Mannheim, Mannheim, Germany

Keywords: stochastic dominance, bayes factors, ordinal scales, meaningful comparisons, Likert items

https://doi.org/10.1525/collabra.38594

Collabra: Psychology

Vol. 8, Issue 1, 2022

Ordinal-scale items—say items that assess agreement with a proposition on an ordinal rating scale from *strongly disagree* to *strongly agree*—are exceedingly popular in psychological research. A common research question concerns the comparison of response distributions on ordinal-scale items across conditions. In this context, there is often a lingering question of whether metric-level descriptions of the results and parametric tests are appropriate. We consider a different problem, perhaps one that supersedes the parametric-vs-nonparametric issue: When is it appropriate to reduce the comparison of two (ordinal) distributions to the comparison of simple summary statistics (e.g., measures of location)? In this paper, we provide a Bayesian modeling approach to help researchers perform meaningful comparisons of two response distributions and draw appropriate inferences from ordinal-scale items. We develop four statistical models that represent possible relationships between two distributions: an unconstrained model representing a complex, non-ordinal relationship, a nonparametric stochastic-dominance model, a parametric shift model, and a null model representing equivalence in distribution. We show how these models can be compared in light of data with Bayes factors and illustrate their usefulness with two real-world examples. We also provide a freely available web applet for researchers who wish to adopt the approach.

It is hard to overstate the popularity of ordinal data in social science research. Applications of ordinal variables such as *Likert items* to assess respondents' opinions, affective states, or unobservable behavior have become customary in political science, economics, educational research, health sciences, and psychology. Likert items refer to statements or questions with discrete, naturally ordered response categories (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018; see also Likert, 1932). A key question in many applications is how responses on Likert items differ between two conditions (say, two groups of respondents). Although there is near universal recognition that Likert items are ordinal variables, these comparisons are commonly characterized with means and t-tests. While some have defended the use of parametric statistics in the context of Likert data (Norman, 2010), others have criticized it as a biased and error-prone practice (Liddell & Kruschke, 2018; Winship & Mare, 1984). The typical recommendation is to rely on nonparametric statistics instead to ensure robust inferences (e.g., Jamieson, 2004; Kuzon et al., 1996; Nanna & Sawilowsky, 1998).

The ordinal-vs-metric issue is well known and there is a large body of literature on it (e.g., Bürkner & Vuorre, 2019; Clason & Dormody, 1994; Jamieson, 2004; Kuzon et al., 1996; Liddell & Kruschke, 2018; McKelvey & Zavoina, 1975; Nanna & Sawilowsky, 1998; Norman, 2010; Sullivan & Artino, 2013; Winship & Mare, 1984). Following Townsend (1990), however, we believe that there is a different, far more fundamental issue that has received considerably less attention (but see Clason & Dormody, 1994): In the usual course of testing the effect of condition on some outcome variable, researchers typically rely on the comparison of summary statistics (e.g., measures of central tendency). These comparisons may establish a certain order relationship at the level of this summary statistic, for example, "The mean value is larger in Condition A than in Condition B". They do not imply, however, that this relationship holds in a more general sense, that is, at the level of distributions. In fact, if the relationship between distributions as a whole is qualitatively different from that between the considered summary statistics, a comparison of the latter would not be meaningful and may even mislead the analyst. This state holds across different levels of measurement, and it is true for parametric and nonparametric tests alike.

Based on Townsend's (1990) theory of hierarchical inference, we argue that when comparing responses on a Likert item between two conditions, researchers should first test for order relationships at the level of distributions. If

[a] martin.schnuerch@uni-mannheim.de

a certain ordering holds at this level, it is also implied at a lower level, that is, for a summary statistic such as the mean or median. The reverse, of course, is not true. An ordering may hold for some summary statistic but not for the distribution as a whole, that is, consideration of a summary statistic in this case does not represent the phenomena of interest. Tests of summary statistics are meaningful in our opinion only when the ordering of the summary statistic indeed represents the ordering of distributions.

This condition where distributions order is called *stochastic dominance*, and it is a well-known concept for example in economics (Abadie, 2002; Levy, 1992). Stochastic dominance describes an order relationship between distributions such that one cumulative distribution function is "greater" (or "less") than the other cumulative function for all possible values (Speckman et al., 2008). Heathcote et al. (2010) developed methods to assess stochastic dominance and compared their performance with that of existing procedures (e.g., Kolmogorov-Smirnov tests). These tests are only suited for continuous data, however, which renders them inappropriate for Likert items. In fact, the limited availability of suitable test procedures may be one of the reasons why stochastic dominance is rarely considered in applications with Likert data (cf. Madden, 2009; Tubeuf & Perronnin, 2008).

In this paper, we provide a Bayes-factor approach to help researchers use data to assess stochastic dominance and draw appropriate inferences from Likert items. In the following, we briefly outline conventional approaches to analyzing Likert items, and highlight the role of stochastic dominance. We then develop four statistical models that represent possible order relationships between two response distributions: An unconstrained model representing a complex, non-ordinal relationship, a nonparametric stochastic-dominance model, a parametric shift model, and a null model representing equivalence in distribution. We show how these models may be evaluated in light of data by means of Bayes factors and present a user-friendly web applet for readers who wish to adopt the analysis in their own research. Finally, we demonstrate the usefulness of the approach by applying it to two real-world examples, and assess the sensitivity of Bayes factor model comparisons to reasonable variations in prior settings.

Likert-Item Distributions

To illustrate why the parametric-vs-nonparametric debate does not address the heart of the problem, consider the following hypothetical example: Suppose we wanted to compare the frequency of being sad between first-year Marines and first-year college students. From each group, we let 100 individuals indicate on a 5-point Likert item how often they felt sad, with response options ranging from "never" to "always". Table 1 shows hypothetical data for two different scenarios labeled plainly Scenario I and Scenario II. For each scenario, we may ask whether there is a difference between Marines and college students.

A nonparametric alternative to t-tests for addressing this question is the Wilcoxon rank-sum test. Unlike the t-test, the Wilcoxon test does not consider the difference between values but only the rank order. Nanna and Sawilowsky (1998) compared the performance of both tests in the context of Likert data and found that the nonparametric test outperformed the parametric test in terms of Type I error control and statistical power. Despite these differences in performance, both procedures have in common that they compare distributions by comparing central tendencies. In Scenario I, the response distributions of Marines and college students differ in their central tendencies. College students seem to be more often sad than Marines, and both a t-test and a Wilcoxon rank-sum test will detect that difference. Importantly, this relationship holds qualitatively across the response scale: College students' reported frequency of being sad is unambiguously higher than that of Marines.

A different picture emerges in Scenario II: Comparing Marines' and college students' answers by means of central tendencies implies the same ordering as in Scenario I, that is, students seem to report being sad more often than Marines. This ordering is not preserved at the level of distributions, however. While many Marines report *never* being sad, many also report *always* being sad. Thus, tests of central tendencies, parametric and nonparametric test alike, do not allow for a meaningful comparison of conditions (Clason & Dormody, 1994).

The crucial difference between the scenarios is that in Scenario I, the distributions are stochastically dominant, whereas in Scenario II, this dominance does not hold. Stochastic dominance describes the relationship among cumulative probabilities, and for observed data, may be visualized using *cumulative proportions*. Table 2 presents these cumulatives for Scenarios I and II. Each number denotes the proportion of people whose response fell into the respective or a lower category. For example, the first two values for Marines in Scenario I are .30 and .55, and these values indicate that 30% of Marines report to be never sad and 55% report to be either never sad or rarely sad. The key property here is the comparison of these cumulatives to those for college students. The values for college students indicate that only 20% are never sad and 40% are either never or rarely sad. The cumulative proportions for the Marines are always at least as great as those for the college students, and this property holds across all categories.

The pattern is more complex in Scenario II. 60% of Marines report to be sometimes, rarely, or never sad, while only 36% of college students do. Thus, for these three categories, Marines report a lower frequency of being sad than college students. This relationship reverses at *Often*, however. While 80% of college students report to be sad often or less, leaving 20% to be always sad, only 70% of Marines chose *Often* or less, leaving 30% for the highest category. There is no stochastic dominance in this case, Marines are both more frequently never sad and more frequently always sad.

Bayesian Models for Ordinal-Scale Data

So far, we focused on cumulative proportions, which are sample-level data. As researchers, however, we are typically interested in the underlying population-level probabilities,

Table 1. Ratings Distributions for Hypothetical Sadness Example.

	Never	Rarely	Sometimes	Often	Always
Scenario I					
Marines	30	25	20	15	10
College Students	20	20	20	20	20
Scenario II					
Marines	40	15	5	10	30
College Students	5	12	19	44	20

Note. Question: How often do you feel sad?

Table 2. Cumulative Proportions for Hypothetical Sadness Example.

	Observed Proportions					
	Never	Rarely	Sometimes	Often	Always	N
Scenario I						
Marines	0.30	0.55	0.75	0.90	1.00	100
College Students	0.20	0.40	0.60	0.80	1.00	100
Scenario II						
Marines	0.40	0.55	0.60	0.70	1.00	100
College Students	0.05	0.17	0.36	0.80	1.00	100

Note. Question: How often do you feel sad?

that is, the behavior of these proportions in the large-sample limit. To assess whether stochastic dominance holds in population, we need a hypothesis test suitable for ordinal data.

Tests of stochastic dominance that assume continuous data (such as the Kolmogorov-Smirnov test) are not appropriate for Likert data. As an extension of one of these tests, Yalonetzky (2013) developed a method for testing stochastic dominance with ordinal data. The test is based on the asymptotic approximation of the multinomial distribution to a multivariate normal distribution. Klugkist et al. (2010) developed a Bayesian hypothesis testing procedure for inequality/equality constrained hypotheses for contingency tables. This nonparametric approach is very general and allows the analyst to test certain expected orderings of cell probabilities. Thus, the method could be used to test a certain ordering of response probabilities implied by stochastic dominance in Likert data. Heck and Davis-Stober (2019) discuss a similar approach for testing order constraints, including stochastic dominance, in multinomial models (see also Sarafoglou et al., 2021).

We suggest a related approach to assessing stochastic dominance with Likert data. Our main goal is to provide four models that encode a series of nested nonparametric and parametric constraints. While the aforementioned methods could also be used to encode and test nonparametric constraints, the approach that we propose makes it straightforward to specify and test both nonparametric and parametric constraints.

Under the most constrained of the four models, distributions across the two conditions are identical. At the next most constrained level, the distributions differ but this difference is captured in a (semi-) parametric model that underlies *ordinal-regression* (also referred to as *ordered-probit* or *cumulative*) model settings (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018; McKelvey & Zavoina, 1975; Winship & Mare, 1984). In the third model, the semi-parametric form is further relaxed, leaving a model that has only a nonparametric stochastic dominance constraint. And finally, even this constraint is relaxed, allowing for more complex, non-ordinal relationships. By comparing the strength of evidence from data for these four models, researchers can make insightful, meaningful comparisons across conditions.

Ordinal-Regression Setup

It is convenient to start with the well-known ordinal-regression approach (McKelvey & Zavoina, 1975; Winship & Mare, 1984). Here, the observed variable (i.e., the choice of a response category) results from the categorization of an underlying continuous variable. Consider a hypothetical survey study where respondents are asked to rate a statement on a 5-point scale ranging from “Strongly Disagree” to “Strongly Agree”. The model posits that agreement with this statement can be represented as a continuous, latent variable. This latent variable maps onto rating categories by partitioning the latent space into regions. These regions are defined by thresholds, and the probability of a response falling into a certain category is simply the area under the latent probability distribution between the respective thresholds (Winship & Mare, 1984). The model setup is illustrated in Figure 1. Note that this setup is conceptually equivalent to that underlying signal-detection theory.

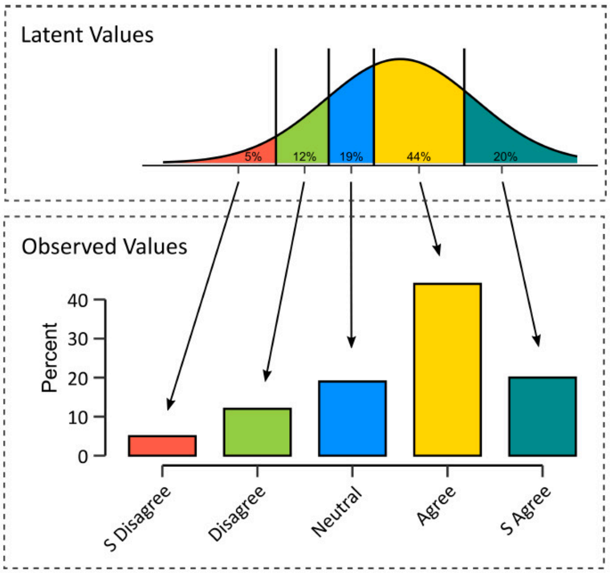


Figure 1. Ordinal-Regression Model

The latent variable is typically assumed to be normally distributed, although the model may be based on other probability distributions (e.g., a logistic function; Bürkner & Vuorre, 2019). The upper panel of Figure 1 shows a latent variable that is partitioned into five regions by four thresholds (represented by the vertical lines). Whenever the latent value exceeds a threshold, the observed response is the associated category (lower panel). Thus, the probability of a latent value falling into a certain region corresponds to the probability of observing the associated response. For more details, we refer the reader to an accessible tutorial by Bürkner and Vuorre (2019), who provide an extensive overview of this and related models for the analysis of Likert items.

In the usual ordinal-regression approach, the thresholds are fixed across conditions and differences in distributions are captured by shifting the central tendency of the latent distribution. This usual approach may be considered *semiparametric* as there is no model on thresholds but a parametric model on the effect of conditions. We are going to start with a fully *nonparametric* model that is an unconstrained generalization of the ordinal-regression approach, and then add in increasing degrees of constraint.

We start by setting the latent distribution for both conditions to a standard normal ($\mu = 0$, $\sigma^2 = 1$). The free parameters in this setup are the category thresholds. Let γ_{ij} denote the threshold between response category j and $j+1$ ($j = 1, \dots, J$) in condition i ($i = 1, 2$). For the setup to be valid, thresholds *within* each condition have to order, that is, $\gamma_{i0} = -\infty \leq \gamma_{i1} \leq \dots \leq \gamma_{iJ} = \infty$. Although it may appear that the choice of identical standard normals is assumptive, in this setup with free threshold parameters, it is not. The latent distribution serves merely as a technical device that maps observed response frequencies onto regions on the real line. Importantly, all observed Likert distributions across conditions may be accounted for by appropriate settings of the thresholds. Thus, at this point, the model is unconstrained, nonparametric, and vacuous; there are as many parameters as degrees of freedom in the data.

To add constraint, it is useful to reparameterize the thresholds as follows:

$$\gamma_{ij} = \alpha_j + x_i \theta_j,$$

where $x_1 = -1/2$ and $x_2 = 1/2$. Here, $\alpha_j = (\gamma_{1j} + \gamma_{2j})/2$ is the average for the jth threshold, and $\theta_j = \gamma_{2j} - \gamma_{1j}$ is the difference for the jth threshold. The key feature of this parameterization is that θ_j denotes a comparison of distributions for the jth threshold. Thus, by placing constraints on θ_j, we can model different types of (ordinal) relationships between the two response distributions.

Models

We specify four statistical models on θ_j, each representing a different constraint on the relationship between conditions. The models are shown in Figure 2, illustrating the construction in the context of our hypothetical sadness example (Table 1).

Unconstrained Model: The first row shows a model that imposes no order constraints on the relationship between conditions. So long as the thresholds order within a condition (which is imposed by the likelihood function), there is no restriction on the values and relative order of thresholds across conditions. We denote this model as $\mathcal{M}_u$, with $\mathcal{M}_u\colon\ \theta_j \in \mathbb{R}$. There are $2 \times (J-1)$ free parameters in this model: $J-1$ mean-threshold parameters (α_j) and equally many difference parameters (θ_j). The unconstrained model can account for any type of relationship between conditions, including complex relationships where response distributions differ in a way that cannot be captured by an order relationship.

Dominance Model: The second row shows the dominance model, $\mathcal{M}_d$. For this model, there are again a total of $2 \times (J-1)$ free parameters. To capture the notion of stochastic dominance, however, we impose an order constraint in this model: $\mathcal{M}_d\colon\ \theta_j \geq 0$. This constraint implies that thresholds are at least as large—and hence, so are cumulative probabilities—in one condition as in the other one. For the example in Figure 2, γ_{11} is the threshold that separates *Rarely* from *Never* for college students, and it has a value of -0.67. Likewise, the value for Marines is denoted γ_{21} and has a value of -0.04. Here, we see that $\gamma_{21} > \gamma_{11}$—Marines have a higher probability of being never sad than college students. Importantly, this inequality holds for all corresponding thresholds, that is, because $\theta_j \geq 0$ for all thresholds it follows that $\gamma_{2j} \geq \gamma_{1j}$ for all threshold pairs.

There are two possible dominance conditions: one in which all $\theta_j \geq 0$ (i.e., $\gamma_{2j} \geq \gamma_{1j}$) and one in which all $\theta_j \leq 0$ (i.e., $\gamma_{2j} \leq \gamma_{1j}$). Whether one or the other or both should be used is a specification decision that researchers should make ahead of time depending on context. We will discuss how these decisions may be made subsequently.

Constant Shift Model: The next row describes a very simple effect where the thresholds in one condition all shift by the same amount compared to the other condition. The model is denoted by $\mathcal{M}_1$, and imposes the parametric constraint $\mathcal{M}_1\colon\ \theta_j = \theta^*$. We include this model because it cor-

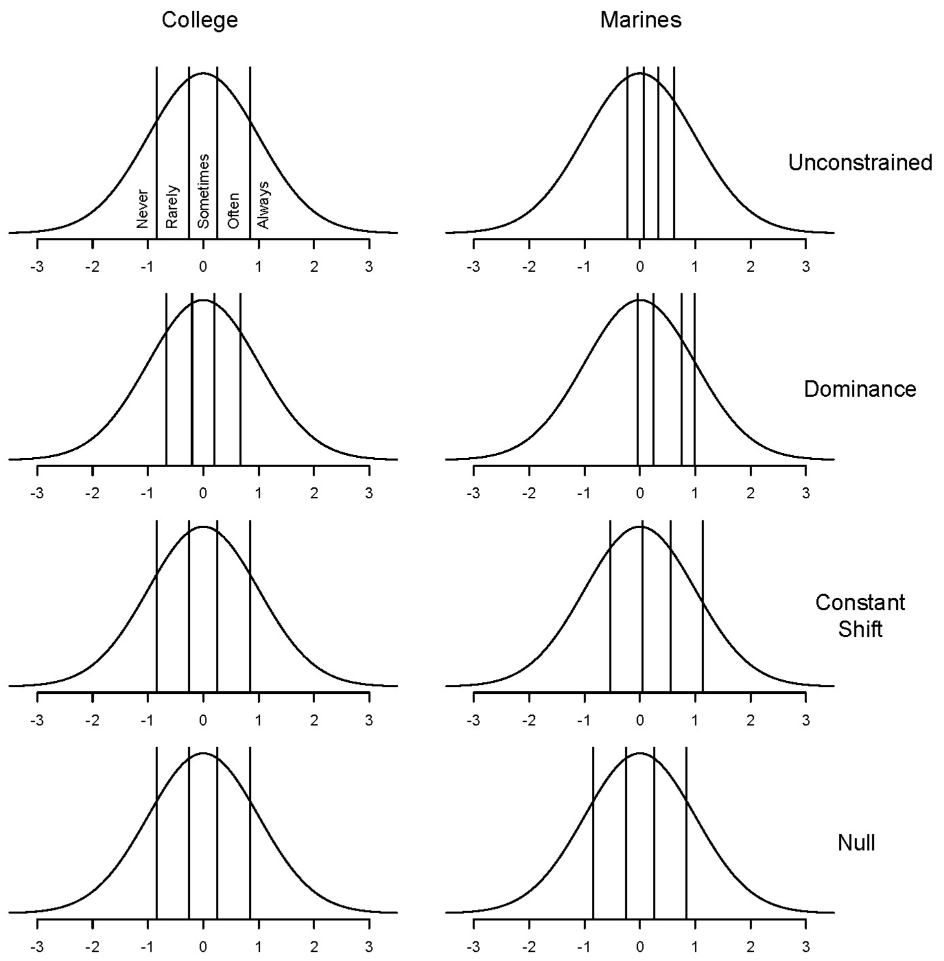


Figure 2. Illustration of Statistical Models

Note. The latent distribution is fixed as a standard normal and latent thresholds are free parameters. The four models, depicted across the rows, capture different types of relationships between conditions (college students vs. Marines).

responds to the classical probit-regression model presented above (see Figure 1). In the probit-regression model, the shifts are in the mean of the normal, but this is mathematically equivalent to fixing the mean and shifting all the thresholds by a constant amount. Unlike the other models we propose in our framework that are nonparametric, the constant shift model imposes a parametric constraint on the latent threshold parameters, that is, constancy is made with respect to the normal distribution. Thus, for this model, the choice of identical latent distributions is indeed a substantive statement about the data. Of note, even though constancy reflects the choice of latent distribution, dominance does not. If thresholds order between conditions for one latent distribution, they must order for all other latent distributions.

In Figure 2, the value of the threshold between *Never* and *Rarely* for Marines is -0.54, and this value is 0.30 greater than the bound between *Never* and *Rarely* for college students. This difference is preserved across corresponding thresholds. For example, the thresholds between *Rarely* and *Sometimes* are 0.05 and -0.25 for Marines and college students, respectively. The difference, 0.30, is the same as between *Never* and *Rarely*. The constant shift model explicitly states that the effect of condition on the ratings can be captured by a single parameter θ^*. It is comprised of J free parameters (i.e., $J-1$ mean thresholds α_j and one difference θ^*). In our view, the constant-shift model is useful for cases where the effect of condition is relatively straightforward and can be captured by a shift in central tendency.

Null Model: The last row depicts the null model which posits that there is no effect of condition. This model is denoted $\mathcal{M}_0$, and imposes the constraint $\mathcal{M}_0\colon\ \theta_j = 0$. Thus, the corresponding thresholds for college students and Marines are identical in this model. For example, the value of the threshold between *Never* and *Rarely* in Figure 2 for college students is -0.84, and this value is the same for the threshold between *Never* and *Rarely* for Marines. Because all the corresponding thresholds are the same in value, the distributions are the same as well. There is no difference among the conditions; hence, there is no effect. The null model has one free parameter for each threshold, that is, $J-1$ parameters in total.

Priors on Parameters

Our approach is Bayesian, and in Bayesian analysis priors are needed on parameters. All four models considered here comprise $J-1$ parameters for the mean thresholds

α_j, so the priors for these parameters should be identical across models. A typical choice for these priors are independent normal distributions (e.g., Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018):

$$\alpha_j \sim \mathrm{Normal}(0, b_\alpha),$$

where b_α is a prior standard deviation setting that must be chosen before analysis.

In contrast to the priors on α_j, the priors on the difference parameters θ_j reflect the substantively motivated constraints under the four models. As for the mean thresholds, we propose a flexible normal distribution as a basis for these priors. Under $\mathcal{M}_u$, we specify independent normal distributions for each θ_j :

$$\theta_j \sim \mathrm{Normal}(0, b_\theta),$$

where b_θ is again specified before analysis. Under $\mathcal{M}_d$, truncated normal distributions are placed on θ_j to impose the notion of stochastic dominance:

$$\theta_j \sim \mathrm{Normal}_T(0, b_\theta),$$

where Normal_T denotes a normal distribution with either an upper or a lower bound at 0, respectively. Under $\mathcal{M}_1$, there is just one difference parameter θ^* and thus,

$$\theta^* \sim \mathrm{Normal}(0, b_\theta).$$

Finally, no prior on θ_j is needed under $\mathcal{M}_0$, as the difference in thresholds between conditions is constrained to be 0.

Before analysis, researchers can adjust the prior parameters b_α and b_θ as needed. Thus, the normal prior setting offers the flexibility to provide substantive context through the choice—and range—of these prior parameters. Here is some guidance for setting b_α and b_θ in practice: Since thresholds are placed on a standard normal, reasonable values of b_α should be around 1.0. Figure 3 shows the marginal prior distribution on mean category probabilities across conditions for 5 rating options and for select values of b_α. For $b_\alpha = 1$, middle panel, the marginal priors have the same distribution, centered around .2, for each of the five rating options. Small values of b_α correspond to a belief that extremes are used excessively at the expense of the middle category (left panel); a large value of b_α corresponds to a belief that extremes are used rarely (right panel). The setting $b_\alpha = 1$ is a good, weakly informative default, and it is hard to imagine reasonable settings smaller than 1/3 and larger than 3.

The prior standard deviation on the difference parameters, in contrast, should typically be much smaller than on the mean thresholds. As for any difference parameter, however, the exact choice depends on the analyst's expectation about how strongly the distributions may differ from each other. Thus, this choice should be determined by substantive, rather than statistical, arguments. For our purposes, we choose a prior standard deviation of $b_\theta = 0.33$, that is, 1/3 of b_α. We address the consequences of this choice and how it affects model comparison results subsequently.

Data Visualization

The four models correspond to the following helpful data visualizations. Much like in signal-detection analysis, the running cumulative proportions become the target for plotting. Table 2 shows the cumulatives for the two hypothetical sadness scenarios. The usual approach is to plot *receiver operating characteristic curves* (ROCs), and an example for Scenario I is shown in Figure 4A. The levels of constraint are as follows: If the null model holds, the ROC curve traces the diagonal. If the shift model holds, then the resulting curve is the stereotypical one (Figure 4A) that is common in memory and perception research. The dominance model implies that the points all lie on one or the other side of the diagonal. The unconstrained model implies only that the points increase on the x and y axes, respectively (Figure 4C). For analyzing real-world contrasts, it is advantageous to plot the differences across the conditions as in Figures 4B and D. The advantage here is that it is easier to spot trends because the y axis may be scaled for differences rather than the entire range from 0 to 1. The constraints now center around the horizontal zero line. The null model corresponds to this line; the shift and dominance model correspond to curves strictly on one side of it; the unconstrained model has no such constraint. Figures 4C and 4D show the ROC and the difference plot for the data in Scenario II.

Bayes Factors

We can measure the strength of evidence from the data for the four models using *Bayes factors* (Jeffreys, 1961), which are a measure of how well each model predicted the data before they are observed (Rouder & Morey, 2018). Readers who are new to Bayes factors are invited to consider one of the many tutorials on their use, and perhaps one of the most helpful resources is the recent 2018 *Psychonomic Bulletin & Review* special issue on Bayesian inference (Vandekerckhove et al., 2018).

There are many approaches to computing Bayes factors. For the models developed here, we use two different approaches as follows: Some models differ in dimensionality. For example, for $J = 5$ response options, there are $2 \times (J - 1) = 8$ parameters in the unconstrained model, $(J - 1) + 1 = 5$ parameters in the shift model, and $J - 1 = 4$ parameters in the null model. Where the models differ by a relatively small number of parameters, we find that the *bridge sampling* approach proposed by Meng and Wong (1996) works well. Gronau et al. (2017) provide a detailed and accessible tutorial on computing Bayes factors with bridge sampling. The approach has been implemented in an R package by Gronau et al. (2020), which we use in our work as well.

We follow a different approach to compare models that have the same number of parameters, namely, the unconstrained and dominance model. The dominance model is more constrained by virtue of the inequalities. Thus, although the models have the same dimensionality, the parameter space for the dominance model is smaller than that for the unconstrained model. In fact, the unconstrained model *encompasses* the dominance model (Heck & Davis-Stober, 2019; Klugkist et al., 2010). When models are encompassed, the Bayes factor may be computed by considering the posterior and prior probabilities of the constraint under the unconstrained model (Gelfand et al., 1992). The

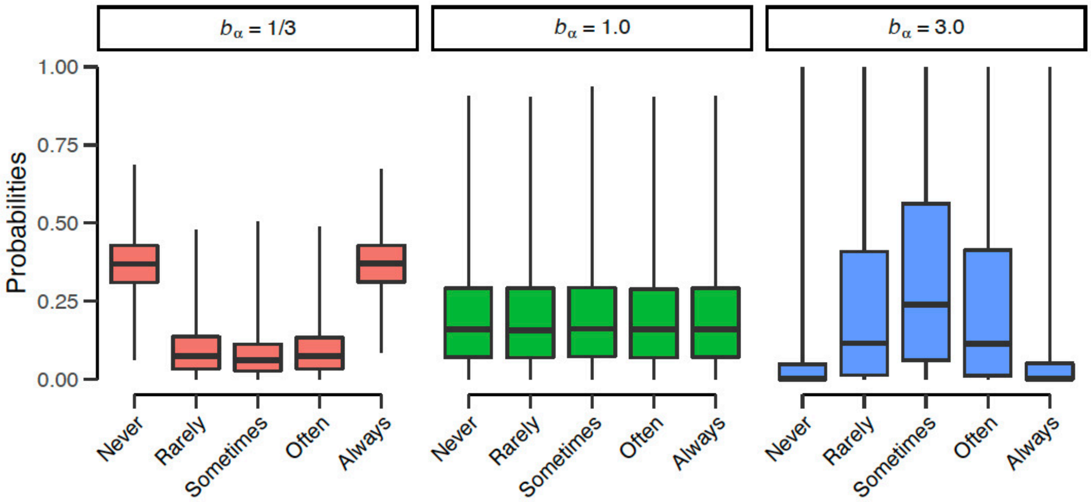


Figure 3. Marginal Prior Distributions on Average Category Probabilities

Note. b_α = Prior standard deviation setting on α_j ; $J = 5$ rating options.

resulting Bayes factor between the dominance and the unconstrained model is

$$B_{du} = \frac{Pr(\mathcal{M}_d|\mathbf{Y})}{Pr(\mathcal{M}_d)}.$$

The first step is calculating the denominator, that is, the prior probability that one distribution dominates another. This calculation may be done by Monte-Carlo simulation from the priors on the collections of α and θ under the unconstrained model. The next step is calculating the numerator. In practice, the computation is surprisingly uncomplicated. We follow the approach discussed in Haaf and Rouder (2017), which is based on the pioneering work of Klugkist et al. (2005). One simply counts the relative frequency of posterior samples under $\mathcal{M}_u$ that satisfy the dominance constraint (see Sarafoglou et al., 2021 for an alternative, efficient routine to calculating Bayes factors for order constraints using bridge sampling). Note that Bayes factors calculated with the encompassing-prior approach are bounded by the prior probability of the constraint under the unconstrained model. Thus, if there is unequivocal evidence that the dominance constraint holds, the Bayes factor may be no larger than $1/Pr(\mathcal{M}_d)$.

As outlined before, there are two dominance conditions because either distribution could possibly dominate the other. A test of stochastic dominance can be two-sided if there is no prediction about which distribution dominates the other. In this two-sided case, the prior probability of stochastic dominance is twice that of a directed test, that is, where a researcher *a priori* predicts that one distribution dominates the other and not the reverse. The posterior probability is estimated as the relative frequency of posterior samples in the predicted direction only. If stochastic dominance is observed in this predicted direction, the corresponding Bayes factor will yield stronger evidence than in the two-sided case. Thus, if theoretical considerations indicate a dominance relation in a specific direction, the Bayes factor should be calculated accordingly.

We do not recommend that researchers compare both stochastic dominance models with one in each direction. This recommendation is a matter of judgment. The motivation is that model comparison and testing should occur when researchers have good reason to suspect an effect in a theoretically meaningful direction. When researchers have no such reasons, exploratory approaches may be more appropriate than model comparison.

Software for Computing Bayes Factors

We created a user-friendly R web applet for analysis. The user inputs the frequency counts in two conditions such as in Table 1. The outputs are Bayes factors for the four models. Additional prior inputs, such as the standard deviations b_α and b_θ may be provided as well. The web applet is available at https://martinschnuerch.shinyapps.io/likertBF/; the underlying source code as well as a set of useful R functions are available at https://github.com/mschnuerch/likertBF.

We illustrate this applet with the example data about sadness in Marines and college students, Scenario I. A screenshot of the applet while analyzing the data is shown in Figure 5. Once the data are inputted, we may press “Plot Data,” and under “Data Visualization,” we may see the diagnostic plots that are shown in Figure 4. Then, to compute Bayes factors, we may press “Start Analysis,” and after some time for sampling, the Bayes factors are returned. We may even choose which dominance model we wish by selecting the respective output option. Let’s say *a priori* we may have thought college students would be more often happy. Because we entered the Marines under Condition 1 and the scale ranges from “never sad” to “always sad”, we specify the one-sided dominance model as “2 > 1”. The results, shown in the center panel, clearly indicate that the constant shift model is preferred. Finally, by clicking “Plot

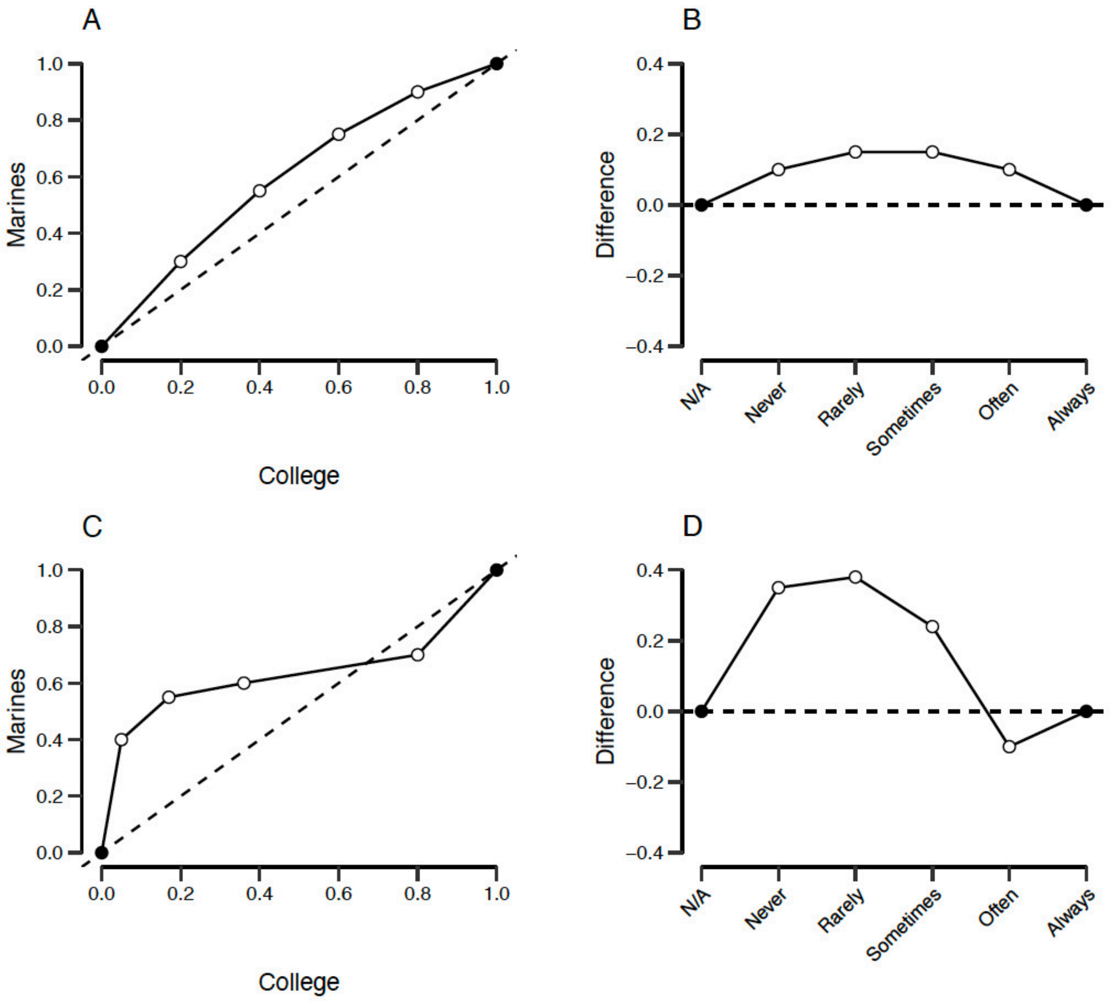


Figure 4. ROC and Difference Plots for Hypothetical Sadness Example

Note. See Table 2. A, B = Scenario I; C, D = Scenario II.

MCMC", we can visually inspect MCMC samples from the unconstrained model for α_j and θ_j.

Applications

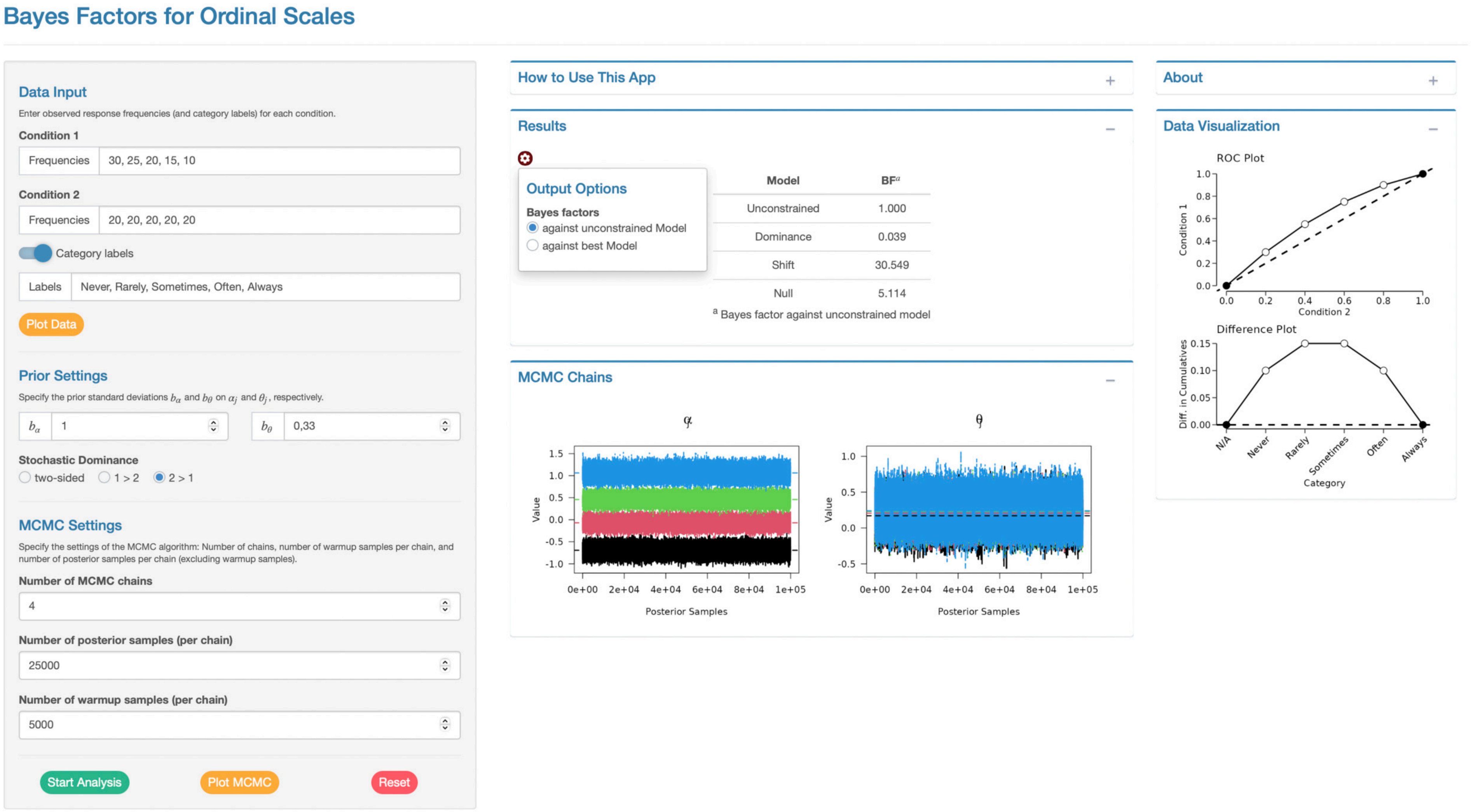
In this section, we provide two real-world examples of these fine-grained analyses. The first example comes from Collingwood et al. (2018) who asked respondents their opinions about controversial policies of the US administration under former president Donald Trump, including the ban on immigration from select Islamic nations and the continuation of the Keystone pipeline project.[1] Collingwood et al. (2018) conducted two survey waves: one when the policy was proposed and the other during implementation. The observed proportions and sample sizes are shown in Table 3.

The second example comes from the Pew Research Center's *Election News Pathways Project* (Pew Research Center, 2020). Over 11, 000 respondents were surveyed about their perception of the Covid-19 pandemic in late March, 2020.[2] We contrast two questions: In one, participants were asked to rate how well US President Trump was responding to the pandemic; in the other, they were asked to rate how well their respective state leaders were responding to the pandemic. The observed proportions and sample sizes are shown in the panel labeled *All* in Table 4.

Collingwood et al. (2018) claimed that the Muslim immigration ban became more popular after it was implemented. We use the four models to assess whether there really was an effect, and if so, whether it may be captured with an order relationship as implied by the dominance and shift models. Figure 6, top left, shows the difference in cumu-

[1] The data set is publicly available from https://github.com/PerceptionAndCognitionLab/bf-likert

[2] The data set is freely available upon registration from https://www.pewresearch.org/politics/dataset/american-trends-panel-wave-64/



Bayes Factors for Ordinal Scales

Data Input

Enter observed response frequencies (and category labels) for each condition.

Condition 1

Frequencies: 30, 25, 20, 15, 10

Condition 2

Frequencies: 20, 20, 20, 20, 20

Category labels

Labels: Never, Rarely, Sometimes, Often, Always

Prior Settings

Specify the prior standard deviations b_α and b_θ on α_j and θ_j, respectively.

b_α: 1

b_θ: 0,33

Stochastic Dominance

○ two-sided ○ 1 > 2 ● 2 > 1

MCMC Settings

Specify the settings of the MCMC algorithm: Number of chains, number of warmup samples per chain, and number of posterior samples per chain (excluding warmup samples).

Number of MCMC chains: 4

Number of posterior samples (per chain): 25000

Number of warmup samples (per chain): 5000

Output Options

Bayes factors

● against unconstrained Model

○ against best Model

Model	BF[a]
Unconstrained	1.000
Dominance	0.039
Shift	30.549
Null	5.114

[a] Bayes factor against unconstrained model

Figure 5. Screenshot of the Accompanying Web Applet

Note. The analyzed data shown in the screenshot correspond to the hypothetical Scenario I in Table 1.

Table 3. Ratings Distributions from Collingwood et al. (2018).

	Observed Proportions					
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N
Immigration Ban[1]						
First Wave	0.30	0.14	0.14	0.14	0.29	411
Second Wave	0.40	0.11	0.09	0.16	0.23	311
Keystone Pipeline[2]						
First Wave	0.42	0.08	0.18	0.13	0.18	409
Second Wave	0.39	0.14	0.12	0.14	0.2	311

Note. 1. Agreement with President Trump's executive order restricting immigration from Syria, Iran, Iraq, Libya, Yemen, Somalia, and Sudan. 2. Agreement with President Trump's executive order allowing for the Keystone and Dakota Access Pipelines.

Table 4. Ratings Distributions from the Election News Pathway Project.

	Observed Proportions				
	Excellent	Good	Fair	Poor	N
All					
Trump	0.24	0.25	0.19	0.32	11491
State Officials	0.21	0.49	0.22	0.08	11432
Democrats					
Trump	0.04	0.14	0.26	0.56	5937
State Officials	0.21	0.48	0.23	0.07	5914
Republicans					
Trump	0.47	0.36	0.11	0.06	5101
State Officials	0.21	0.52	0.2	0.08	5076

Note. How would you rate the job each of the following is doing responding to the coronavirus outbreak? A. Donald Trump. B. Your elected state officials.

latives. As can be seen, the curve does not cross the zero-line, indicating the plausibility of stochastic dominance. The Bayes factors for the four models are shown in Table 5. As expected, the winning model is the one-sided dominance model, followed by the shift model. Hence, we conclude that there is evidence for an effect. The effect is simple and can be reduced to an order relationship. The same analysis may be applied to the question about the Keystone pipeline. For these data, the null has a Bayes factor of at least 2.5-to-1 against any competitor indicating anecdotal evidence for a lack of an effect of wave on the ratings distribution.

Perhaps the most interesting data are those about leadership in the Covid-19 pandemic. Here, we have strong evidence for an indominant effect. The unconstrained model is preferred by several hundred orders of magnitude to any competitor. Donald Trump seems to be a polarizing figure compared to state leaders. People were more likely to give Donald Trump extreme ratings than state leaders. This polarization may be seen in the difference plot in Figure 6 (bottom right panel). Here, the curve crosses zero, and though the deflection may appear slight, it is highly evidential because the sample sizes are so large. Accordingly, it makes little sense to discuss whether Donald Trump is viewed as having responded better or worse than local leaders.

The complexity of the effect is easily resolved in this case by conditioning the data on political-party preference. Among those that are Republican, Donald Trump is judged quite well in responding to the crisis; among those that are Democratic, he is judged quite poorly. This partisan divide is not present among state leaders. Thus, when we condition responses on political-party preference, the dominance model in the expected direction wins.

Sensitivity To Prior Settings

The Bayesian analysis presented here requires the analyst to set the prior standard deviations on mean bounds and effects (b_α, b_θ). Such requirements have given some researchers pause in adopting Bayesian methods. It seems reasonable as a starting point to require that if two researchers run the same experiment and obtain the same data, they should reach similar, if not the same, conclusions. To harmonize Bayesian inference with this starting point, some analysts actively seek to minimize these effects by choosing likelihoods, prior parametric forms, and heuristic methods of inference so that variation in prior settings have minimal influence (Aitkin, 1991; Gelman et al., 2004; Kruschke, 2012; Spiegelhalter et al., 2002).

We reject the starting point above including the view that minimization of prior effects is necessary. The choice of prior settings is important because it affects the models'

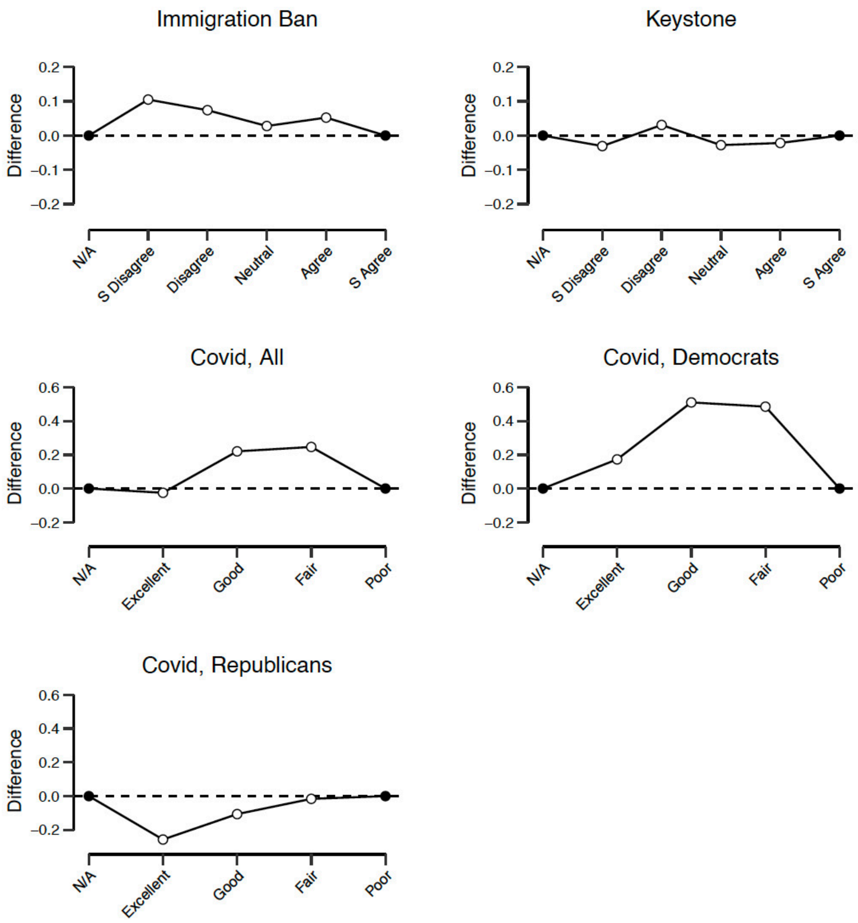


Figure 6. Difference Plots for the Real-World Data in Tables 3 and 4.

Note. There is anecdotal evidence for a constant shift in the top-left panel and for a lack of an effect in the top-right panel. In the middle-left panel, there is strong evidence for an indominant effect, while there is strong evidence for stochastic dominance in the remaining figures.

Table 5. Bayes Factors for Empirical Examples.

	Null	Shift	Dominance	Unconstrained
Immigration Ban	0.21	0.76	1.00	0.13
Keystone Pipeline	1.00	0.29	0.29	0.40
Covid, All	0.00	0.00	0.00	1.00
Covid, Democrats	0.00	0.00	1.00	0.15
Covid, Republicans	0.00	0.00	1.00	0.14

Note. The winning model is assigned a value of 1.00. Bayes factors for all other models are relative to this winning model.

predictions about data. Therefore, these settings necessarily affect model comparison. Whatever this effect, it is the degree resulting from the usage of Bayes rule, which in turn mandates that evidence for competing models is the degree to which they improve predictive accuracy.

When different researchers use different priors, they may arrive at different opinions about the data. This variation is not problematic, however, so long as various prior settings are *justifiable*: The variation in results reflects the legitimate diversity of opinion (Rouder et al., 2016). When different reasonable prior settings suggest conflicting conclusions, the data simply do not afford the precision to arrive at a clear verdict between the positions.

With this argument as context, we may assess whether reasonable variation in prior settings affect Bayes factor conclusions among the models. In Figure 3, we show that $b_\alpha = 1$ is a good default choice for the prior on α_j, and this choice may be made without undue influence on model comparison results. The prior choice on the difference parameters θ_j is more consequential. For the previous analysis, we specified $b_\theta = 1/3$. For this setting, we consider a range from $1/6$ ($1/2$ the original setting) to $2/3$ (2 times the original setting) to be reasonable. Values of $b_\theta < 0.17$

Table 6. Bayes Factors for Modified Election News Pathways Project Data.

	Null	Shift	Dominance	Unconstrained
Covid, All	0.00	0.00	0.13	1.00
Covid, Republicans	0.00	0.00	1.00	0.23

Note. The winning model is assigned a value of 1.00. Bayes factors for all other models are relative to this winning model.

place excessive weight on extremely small differences between conditions, while values of $b_\theta > 0.67$ place excessive weight on overwhelmingly large differences.

To see how variation in this prior setting affects the Bayes factors, we use a modified version of the Election News Pathways Project data. Unfortunately, with $11,000$ observations, the sample size is quite large to be typical of psychological data. A more typical set would have fewer observations, and so for the purposes here we took the frequencies in Table 4 and divided them by 10. We used the complete data set with both Republicans and Democrats because here we found strong evidence for an indominant effect. Along with these data, we used the subset of Republicans as these showed evidence for a simpler structure, namely, stochastic dominance.

The Bayes factors for the modified data set with the same prior setting as in the previous analysis ($b_\alpha = 1.0$ and $b_\theta = 1/3$) are shown in Table 6. Without considering political-party preference, the unconstrained model is still preferred over the others. The closest competitor is the dominance model, and the corresponding Bayes factor is approximately 8-to-1. The reason this value is more moderate than that in Table 5 reflects the reduced sample size. Among Republicans, the dominance model is again preferred over the unconstrained model by a factor of approximately 4-to-1. The question is whether these two values depend heavily on the range of prior settings.

The dependence is shown in Figure 7. Here, Bayes factors of all models against the preferred model within three orders of magnitude (10^{-3}) are displayed. Although the exact figures vary slightly, there is no consequential dependence of Bayes factors across the reasonable range of prior settings. Both for the complete set (left panel) and the subset (right panel), the winning model is preferred over its nearest competitor by a relatively constant amount. Hence, the Bayes factor method provides for evidence that is fairly robust to reasonable variation in prior expectations about data.

Conclusion

Although the use of Likert items is exceedingly popular, we argue here that researchers have overlooked a defining primitive in analysis (Townsend, 1990). Instead of debating the use of parametric vs. nonparametric statistics, we should assess whether or not two response distributions can be meaningfully compared by means of their central tendencies. If there is no order relationship at the level of distributions (i.e., no stochastic dominance), common parametric and even nonparametric tests of differences miss the underlying structure and may mislead the analyst (Clason & Dormody, 1994).

The statistical models developed herein allow for a more fine-grained analysis of Likert and other ordinal-scale items. The null, constant shift, dominance, and unconstrained models provide for a rich description of possible structure in the relationship between two distributions, and strength of evidence from data for them may be stated via Bayes factors. The models as well as the Bayes factor comparisons are straightforward and computationally convenient. We demonstrated their usefulness with two real-world examples and created an easily accessible, user-friendly web applet for researchers.

Although we think that researchers will benefit from the development presented herein, there are also limitations: 1. The concept of the threshold here is not psychological and should not be interpreted as such. In this framework, thresholds describe the proportion of people that endorse particular responses. They do not describe the internal process by which people respond to Likert items. Likewise, the models do not address whether people use the same processes or the same response styles. In this regard, the model is a statistical account for addressing constraints at the population level. 2. Although the unconstrained, dominance, and null models are nonparametric, the constant shift model, which we suspect will be a simple, parsimonious account of condition effects, is parametric. Whether shifts are constant or not depends on the distributional form, and, here, the choice of identical normal distributions for all respondents is a substantive assumption. 3. So far, the development only applies to the comparison of two independent distributions. Of course, psychologists are often interested in more complicated designs. For example, the data from Collingwood et al. (2018) are panel data in which the same people answered in both waves. We do not take into account any shared variation from the panel design. 4. Finally, analysis is not always run for a single item across just two levels of a covariate. It is more typical to use multiple items to construct latent Likert scales. And in this case, questions about a shift or stochastic dominance in the data should be addressed at the scale level.

It is one of the strengths of the proposed analysis framework that it affords the flexibility to incorporate other types of constraints and data structures. In this paper, we focused on the common case of comparing two independent response distributions on a single Likert item. However, future efforts may be devoted to extending our approach to formulate and test other types of constraints across more than two conditions and with multiple items (i.e., Likert scales). At this point, our development constitutes only

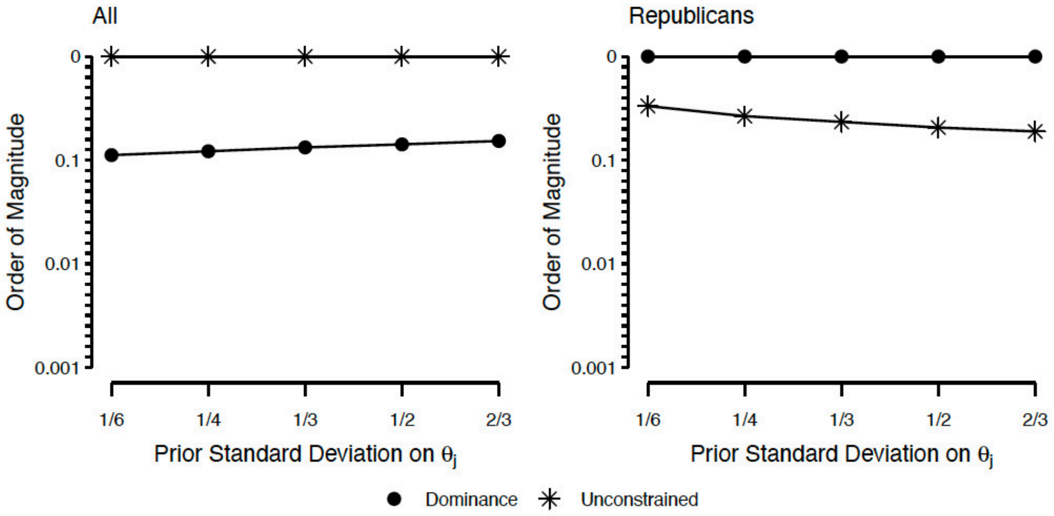


Figure 7. Dependence of Bayes Factors on Prior Settings (b_θ)

Note. Sensitivity analyses were performed on 1/10th of the Election News Pathways Project data. Only models with Bayes factors within three orders of magnitude (10^{-3}) against the preferred model are shown.

a useful first step toward a more complete framework of meaningful analysis of ordinal-scale items.

Contributions

The first author contributed to the theoretical development and implementation of the models and the bridge sampling routine, developed and implemented the web applet, acquired and analyzed the data, and drafted and revised the manuscript; the second author contributed to the theoretical development of the models and revised the manuscript; the third author contributed to the theoretical development of the models and revised the manuscript; the last author contributed to the theoretical development and implementation of the models and the bridge sampling routine, acquired and analyzed the data, and drafted and revised the manuscript.

Competing Interests

The authors declare no competing interests. The second author is an associate Editor at *Collabra: Psychology*.

Funding Information

The first and the last author were supported by a grant from the German Research Foundation to the research training group Statistical Modeling in Psychology (GRK 2277); the third author was supported by a Van der Gaag Fund, Royal Netherlands Academy of Arts & Sciences.

Data Accessibility Statement

The data used in the first real-world application example by Collingwood et al. (2018) are available from https://github.com/PerceptionAndCognitionLab/bf-likert. The data for the second example by Pew Research Center (2020) are freely available upon registration from https://www.pewresearch.org/politics/dataset/american-trends-panel-wave-64/. All code for analyses and figures is included in the R Markdown file of this manuscript. The Markdown file and supporting files are curated at https://github.com/PerceptionAndCognitionLab/bf-likert.

Submitted: April 20, 2021 PDT, Accepted: August 30, 2022 PDT

References

Abadie, A. (2002). Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *Journal of the American Statistical Association*, *97*(457), 284–292. https://doi.org/10.1198/016214502753479419

Aitkin, M. (1991). Posterior Bayes Factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(1), 111–128. https://doi.org/10.1111/j.2517-6161.1991.tb01812.x

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Clason, D., & Dormody, T. (1994). Analyzing Data Measured by Individual Likert-Type Items. *Journal of Agricultural Education*, *35*(4). https://doi.org/10.5032/jae.1994.04031

Collingwood, L., Lajevardi, N., & Oskooii, K. A. R. (2018). A Change of Heart? Why Individual-Level Public Opinion Shifted Against Trump's "Muslim Ban." *Political Behavior*, *40*(4), 1035–1072. https://doi.org/10.1007/s11109-017-9439-z

Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *Journal of the American Statistical Association*, *87*(418), 523–532. https://doi.org/10.1080/01621459.1992.10475235

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Chapman and Hall.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). Bridgesampling: An *R* Package for Estimating Normalizing Constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Haaf, J. M., & Rouder, J. N. (2017). Developing Constraint in Bayesian Mixed Models. *Psychological Methods*, *22*(4), 779–798. https://doi.org/10.1037/met0000156

Heathcote, A., Brown, S., Wagenmakers, E. J., & Eidels, A. (2010). Distribution-Free Tests of Stochastic Dominance for Small Samples. *Journal of Mathematical Psychology*, *54*(5), 454–463. https://doi.org/10.1016/j.jmp.2010.06.005

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87. https://doi.org/10.1016/j.jmp.2019.03.004

Jamieson, S. (2004). Likert Scales: How to (Ab)Use Them. *Medical Education*, *38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

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

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69. https://doi.org/10.1111/j.1467-9574.2005.00279.x

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian Evaluation of Inequality and Equality Constrained Hypotheses for Contingency Tables. *Psychological Methods*, *15*(3), 281–299. https://doi.org/10.1037/a0020137

Kruschke, J. K. (2012). Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology: General*, *142*, 573–603. https://doi.org/10.1037/e502412013-055

Kuzon, W. M. J., Urbanchek, M. G., & McCabe, S. (1996). The Seven Deadly Sins of Statistical Analysis. *Annals of Plastic Surgery*, *37*(3), 265–272. https://doi.org/10.1097/00000637-199609000-00006

Levy, H. (1992). Stochastic Dominance and Expected Utility: Survey and Analysis. *Management Science*, *38*(4), 555–593. https://doi.org/10.1287/mnsc.38.4.555

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *22*(140), 5–55.

Madden, D. (2009). Mental stress in Ireland, 1994-2000: A stochastic dominance approach. *Health Economics*, *18*(10), 1202–1217. https://doi.org/10.1002/hec.1425

McKelvey, R. D., & Zavoina, W. (1975). A Statistical Model for the Analysis of Ordinal Level Dependent Variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/10.1080/0022250x.1975.9989847

Meng, X., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, *3*(1), 55–67. https://doi.org/10.1037/1082-989x.3.1.55

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Pew Research Center. (2020, March 26). *Worries about coronavirus surge, as most Americans expect a recession – or worse*. https://www.pewresearch.org/politics/2020/03/26/worries-about-coronavirus-surge-as-most-americans-expect-a-recession-or-worse/

Rouder, J. N., & Morey, R. D. (2018). Teaching Bayes' Theorem: Strength of Evidence as Predictive Accuracy. *The American Statistician*, *73*(2), 186–190. https://doi.org/10.1080/00031305.2017.1341334

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra: Psychology*, *2*(1), 6. https://doi.org/10.1525/collabra.28

Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E.-J., & Marsman, M. (2021). Evaluating multinomial order restrictions with bridge sampling. *Psychological Methods*. https://doi.org/10.1037/met0000411

Speckman, P. L., Rouder, J. N., Morey, R. D., & Pratte, M. S. (2008). Delta plots and coherent distribution ordering. *The American Statistician*, *62*(3), 262–266. https://doi.org/10.1198/000313008x333493

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Sullivan, G. M., & Artino, A. R. J. (2013). Analyzing and interpreting data From Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/jgme-5-4-18

Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin*, *108*(3), 551–567. https://doi.org/10.1037/0033-2909.108.3.551

Tubeuf, S., & Perronnin, M. (2008). *New prospects in the analysis of inequalities in health: A measurement of health encompassing several dimensions of health* (p. 48) [Tech. rep.]. University of York.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. https://doi.org/10.3758/s13423-018-1443-8

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, *49*(4), 512. https://doi.org/10.2307/2095465

Yalonetzky, G. (2013). Stochastic dominance with ordinal variables: Conditions and a test. *Econometric Reviews*, *32*(1), 126–163. https://doi.org/10.1080/07474938.2012.690653

Supplementary Materials

Response letter version 3

Download: https://collabra.scholasticahq.com/article/38594-meaningful-comparisons-with-ordinal-scale-items/attachment/101003.pdf?auth_token=qavCoVfW8Qa0T9mzCEFa

Response letter version 2

Download: https://collabra.scholasticahq.com/article/38594-meaningful-comparisons-with-ordinal-scale-items/attachment/101004.pdf?auth_token=qavCoVfW8Qa0T9mzCEFa

Peer Review History

Download: https://collabra.scholasticahq.com/article/38594-meaningful-comparisons-with-ordinal-scale-items/attachment/101005.docx?auth_token=qavCoVfW8Qa0T9mzCEFa