March 29, 2022

Dear Dr. van Ravenzwaaij,

We want to thank you and the reviewers for the helpful feedback and the opportunity to resubmit a revised version of our manuscript. We apologize for the delay in resubmission. It is due to both personal reasons and the fact that we substantially revised the manuscript to address all points raised. Below, you find our detailed responses to your and the reviewers' comments, as well as a description of the changes we made in the revised manuscript. We believe that the revisions have improved the manuscript, and we hope that you will further consider it for publication.

Sincerely,
The authors

Responses to the EDITOR:

We thank the editor for the positive evaluation of our manuscript and the chance to resubmit a revision.

E-1: The editor notes a point from Reviewer 1 about the overall goal and asks, "why it is appropriate to draw inferences about quantities from ordinal responses." As part of addressing this issue, the editor requests we address the implications of assuming identical latent normal distributions. One concern is about the appropriateness of a parametric assumption with ordinal data. **Response:** It appears that Reviewer 1 has misunderstood the mathematical basis of the model. The model has free threshold parameters. Consequently, any continuous distribution may be used to model ordinal responses without any loss of generality. We could have used a uniform or a lognormal or any other distribution. With threshold parameters that are free to vary, any pattern of Likert-response distributions for two conditions may be represented. Hence, there is no substantive assumption underlying the normal distribution in our model. Through the free threshold parameters, the model is nonparametric. The only time the normal comes into play as a substantive constraint is in the shift model. Here, the shift is defined with reference to the normal, and the distribution plays an important role. **Revisions**: We substantially rewrote and restructured the introduction and the presentation of our modeling strategy. We now clearly communicate the central problem that we consider (pp. 3–5) and how it can be addressed with the proposed approach (pp. 10–16). We also explain the ordinal-regression setup that is central to our approach and its underlying assumptions more thoroughly (p. 9). Regarding the role of the normal, we write on p. 10: "*We start by setting the latent distribution to a standard normal (mu = 0, sigma^2 = 1). Note that this does not reflect a substantive assumption about the level of measurement or the distribution of the latent construct. In this modeling setup, the latent distribution serves merely as a technical device that maps observed response frequencies onto regions on the real line.*"

E-2: The editor writes, "both reviewers request a more thorough discussion of previous existing literature". **Response:** We thank the reviewers for pointing our attention to relevant literature. We agree that the manuscript benefits from a more thorough discussion of related approaches. **Revisions:** We followed the suggestions and included the suggested references in the revised manuscript (p. 8). Additionally, we more thoroughly embedded our approach in the literature on stochastic dominance and Townsend's (1990) inference hierarchy (see pp. 4 and 8).

Responses to REVIEWER 1

We thank Reviewer 1 for helpful comments on our manuscript.

R1-1: The reviewer expresses concern that "the assumption of normal distributions for latent variables" is ungrounded and arbitrary. **Response & Revisions:** Please see point E-1 above.

R1-2: The reviewer criticizes that the proposed analytic framework does not provide meaningful inferences because it cannot distinguish between differences in latent constructs and in response styles. **Response:** We fully agree with the reviewer's assessment that this distinction is not possible within the proposed framework. However, we never claimed that it was possible, nor is it something that we want to achieve with this analytic approach. In fact, the situation that we consider does not allow for such an analysis. As each person provides only a single response, it is impossible to decompose response behavior into response styles and "true" scores. We think in many regards, this remark of the reviewer is idiosyncratic and may reflect their current research question rather than ours. **Revisions:** We now explain our goals more clearly to minimize the probability of this misconstrual (see pp. 3–7). The central point of our critique of current practice is not about response styles, levels of measurement, or the use of parametric tests on ordinal data, but rather about reducing the comparison of two (ordinal) distributions to the comparison of simple summary statistics (e.g., central tendencies). The problem that we consider is related to Townsend's (1990) proposed "inference hierarchy": We argue that an order relationship between two summary statistics does not imply the same relationship at the level of distributions. Thus, if we test (parametrically or nonparametrically) a relationship between two summary statistics that does not hold across all values of the scale, the test may not be meaningful. It is in this sense that our approach provides for meaningful comparisons, because it tests (order) relationships at the level of distributions. This is our research question; it is not about response styles.

R1-3: The reviewer criticizes that although we masked our names, the manuscript nonetheless contains information that makes it possible to identify authors' identities. Regrettably, the reviewer further states that we openly defied journal policies and should consider a different journal. **Response:** While we appreciate the reviewer's constructive criticisms, the tone and substance of this comment are not appreciated. In our cover letter, we alerted the editor to the identifying information in case any further remediation was needed. We trust the editor's judgment in this regard.

R1-4: The reviewer points out a few minor infelicities. **Response:** We thank the reviewer for these helpful remarks. **Revisions:** We implemented the suggested changes and replaced "relation" with "(order) relationship" and "bound" with "threshold". Furthermore, we thoroughly checked and revised the manuscript with respect to inconsistencies and ambiguities.

R1-5: The reviewer notes that the exemplary category labels in the Shinyapp are confusing because they may be mistaken for default values rather than examples. **Response and Revisions:** We adjusted the Shinyapp to avoid this potential source of confusion.

Responses to REVIEWER 2

We thank Reviewer 2 for their positive evaluation of our manuscript and for insightful comments and suggestions.

R2-1: The reviewer suggests rewriting the introduction from a broader perspective that is better placed in literature on statistical methodology. **Response and Revisions:** As per the reviewer's suggestion, we substantially rewrote the introduction. In the revised manuscript, we state the problem that we address more clearly (i.e., assessing order relationships at the level of distributions) and we provide a better embedding for our approach in existing statistical literature. Please see pp. 3–5.

R2-2: The reviewer asks for a more comprehensive overview of related methods that address the considered problem beyond t-tests, and mentions the method developed by Klugkist et al. (2010) for analysis of contingency tables. **Response:** We thank the reviewer for this suggestion and for pointing our attention to the work by Klugkist et al. (2010). Indeed, their approach allows for the assessment of stochastic dominance among two response distributions on a Likert item. The approach can be considered fully nonparametric because a Dirichlet prior is put directly on cell probabilities. Thus, it could be used to test the nonparametric unconstrained, dominance, and null models. However, it would not allow the analyst to specify a parametric shift model. We think that this model is useful to identify cases where the effect of condition can be captured by a shift in central tendency, which is why we prefer the ordinal-regression setting that we propose in the manuscript. **Revisions:** The revised manuscript now includes a paragraph on related methods to assess stochastic dominance with ordinal data. Among others, we discuss Klugkist et al.'s (2010) approach and how it relates to our method. Please see p. 8.

R2-3: The reviewer provides four enumerated critiques of our prior specification in the proposed approach. We outline the critiques below and provide our responses in turn:

> **R2-3.1:** Priors should be better motivated due to Bayes factors' sensitivity to prior settings. **Response and Revisions:** We believe that prior specification should generally be motivated by substantive arguments, and we provide guidance on how substantive considerations affect prior specification in the manuscript. The normal prior is a powerful device to express substantive belief (see our response to R2-3.2 below). However, we fully agree that sensitivity to prior choices should be considered, and we do so on pp. 27–28 (see also Figure 7). We believe that addressing the issue head-on is one of the strengths of our manuscript.
>
> **R2-3.2:** Many alternative choices for the prior could be considered. **Response and Revisions:** We fully agree with the reviewer. It is important to note, however, that the current proposal is not one prior but a family of priors that can be tuned as needed by adjusting b_α and b_θ. Thereby, substantive researchers can provide substantive context through the choice (and, in the context of a sensitivity analysis, the range) of these setting. The normal distribution is a widely used prior in the context of ordinal-regression models (e.g., Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018), and we provide guidance for the choice of b_α and b_θ by discussing the substantive implications of certain choices and visualizing the marginal prior distributions on average category probabilities. Please see pp. 14–16 and Figure 3 for discussion.
>
> **R2-3.3:** It would be helpful to give the exact priors under all four models explicitly. **Response and Revisions:** In the revised manuscript, we provide the exact priors on θ_j under

the four models and illustrate how they relate to the substantively motivated constraints of these models. Please see p. 15.

R2-3.4: Normal priors might have thinner tails than the likelihood, possibly resulting in information inconsistency. **Response:** We thank the reviewer for pointing out this interesting issue. For the linear model where data have full support on the real line, this is a nuanced point. Somewhat counterintuitively, it has been shown that normal priors on location parameters in the linear model have a funny problem that, while noteworthy from an academic perspective, has no effect in practice. The problem occurs when the data are small in number and extreme in value. For example, we wish to know whether a true mean is zero and we observe a sample mean of 10 billion with 10 pieces of data. As that sample mean approaches infinity, the Bayes factor does not approach infinity for the alternative, but asymptotes to an exceedingly high value. The problem is well known and remedied in two ways. One is to use a thicker tail, say a Cauchy on the location parameters (this approach is used in Rouder/Morey default Bayes factors). The other is to assume that the variance is known (as in BIC). After careful consideration, we are quite comfortable with the priors chosen here for the following reason. We are using not a linear model but a generalized linear model with a link function that compresses the extremes. With our link, there is no free scale parameter as the data are categorical (multinomial likelihood). The case therefore is more analogous to the known-variance case, the case where there is no inconsistency. Our own thought is that it would take a dissertation's worth of analysis to possibly figure out the theoretical ramifications of thin vs thick tails. Importantly, we think there are no practical issues. Thus, addressing these subtleties is tangential for the purposes of our article. **Revision**s: None.

R2-4: The reviewer argues that one could use virtually flat priors for testing one-sided order models. **Response:** Yes, we agree. But we strongly argue against choosing priors based on comparisons in this way. Instead, we prefer a consistent set of prior specifications that may be used on all similar models. One of the advantages of consistency and disadvantages of the suggested alternative is in regard to the transitivity of Bayes factors: If we change the unconstrained model for different comparisons, say null and one-sided dominance, then there is no more transitivity of the Bayes factors across models. We think the transitivity is too important from a philosophical perspective and worry about a loss of interpretability across the set of specified models. **Revisions**: None.

R2-5: The reviewer notes that the paper focuses on four specific models for comparing two conditions, although the method could be used to test other types of constraints and multiple conditions. **Response:** We fully agree with the reviewer. Indeed, we consider it one of the strengths of the proposed approach that many more extensions and applications are possible. However, we feel that considering additional types of constraints or multiple group comparisons would inflate the paper. Thus, to make our point and introduce the proposed analytic framework, we explicitly decided to focus on the four considered models. **Revisions:** We now write on p. 30: "*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 a useful first step toward a more complete framework of meaningful analysis of ordinal-scale items.* "

R2-6: The reviewer notes that the unconstrained model does not necessarily violate order constraints, which we stated on p. 9 of the original manuscript, and recommends a different wording. **Response:** We thank the reviewer for pointing our attention to this inaccuracy. **Revisions:** We restructured the "Models" section to present the models in ascending order of restrictiveness (pp. 11–14). We first present the unconstrained model that imposes no order constraints (although it of course permits an order relationship). Subsequently, we present the other models that are nested within the unconstrained model by virtue of additional (order) constraints on θ_j.

R2-7: The reviewer recommends including the complement model, instead of the unconstrained model, to test the dominance model. **Response and Revisions**: We disagree with this point. As we now write on p. 20, "*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 reason, exploratory approaches may be more appropriate than model comparison*".

R2-8: The reviewer notes that referring to an unrefereed blog post is not recommended. **Response:** We agree with the reviewer that one should generally be cautious when including references to non-peer-reviewed sources. In this particular case, we discovered by simulation that the prior probability of stochastic dominance can be calculated as P(M_d) = 2/J. The mathematical proof that we refer to was published as a blog post in response to this discovery. We have checked the proof to the best of our abilities, and we know of no cases where the proof doesn't hold. **Revisions:** To address the reviewer's concern, we revised the reference in question (p. 19). We now state that it can be shown that P(M_d) = 2/J, and we explain in a footnote that the result was first discovered by simulation and later corroborated by a non-peer-reviewed mathematical proof.