Journal of Experimental Social Psychology 79 (2018) 328–348

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Analyzing ordinal data with metric models: What could possibly go wrong?☆

Torrin M. Liddell*, John K. Kruschke

Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington, IN 47405-7007, United States of America

ARTICLE INFO

Handling editor: Roger Giner-Sorolla

Keywords:
Ordinal data
Likert
Ordered-probit
Bayesian analysis

ABSTRACT

We surveyed all articles in the *Journal of Personality and Social Psychology* (JPSP), *Psychological Science* (PS), and the *Journal of Experimental Psychology: General* (JEP:G) that mentioned the term "Likert," and found that 100% of the articles that analyzed ordinal data did so using a metric model. We present novel evidence that analyzing ordinal data as if they were metric can systematically lead to errors. We demonstrate false alarms (i.e., detecting an effect where none exists, Type I errors) and failures to detect effects (i.e., loss of power, Type II errors). We demonstrate systematic *inversions* of effects, for which treating ordinal data as metric indicates the opposite ordering of means than the true ordering of means. We show the same problems — false alarms, misses, and inversions — for interactions in factorial designs and for trend analyses in regression. We demonstrate that averaging across multiple ordinal measurements does not solve or even ameliorate these problems. A central contribution is a graphical explanation of how and when the misrepresentations occur. Moreover, we point out that there is no sure-fire way to detect these problems by treating the ordinal values as metric, and instead we advocate use of ordered-probit models (or similar) because they will better describe the data. Finally, although frequentist approaches to some ordered-probit models are available, we use Bayesian methods because of their flexibility in specifying models and their richness and accuracy in providing parameter estimates. An R script is provided for running an analysis that compares ordered-probit and metric models.

1. Introduction

Ordinal data are often analyzed as if they were metric. This common practice has been very controversial, with staunch defenders and detractors. In this article we present novel evidence that analyzing ordinal data as if they were metric can systematically lead to errors. We demonstrate false alarms (i.e., detecting an effect where none exists, Type I errors) and failures to detect effects (i.e., loss of power, Type II errors). We demonstrate systematic *inversions* of effects, for which treating ordinal data as metric indicates the opposite ordering of means than the true ordering of means. We show the same problems — false alarms, misses, and inversions — for interactions in factorial designs and for trend analyses in regression. We demonstrate that averaging across multiple ordinal measurements does not solve or even ameliorate these problems. We provide simple graphical explanations of why these mistakes occur. We point out that there is no sure-fire way to detect these problems by treating the ordinal values as metric, and instead advocate use of ordered-probit models (or similar) because they will better describe the data. Finally, although frequentist approaches to some ordered-probit models are available, we use Bayesian methods because of their flexibility in specifying models and their richness and accuracy in providing parameter estimates.

2. Ordinal data and approaches to modeling them

Ordinal data commonly occur in many domains including psychology, education, medicine, economics, consumer choice, and many others (e.g., Carifio & Perla, 2007; Clason & Dormody, 1994; Feldman & Audretsch, 1999; Hui & Bateson, 1991; Jamieson, 2004; Spranca, Minsk, & Baron, 1991; Vickers, 1999). The ubiquity of ordinal data is due in large part to the widespread use of Likert-style response items (Likert, 1932). A Likert item typically refers to a single question for which the response is indicated on a discrete ordered scale ranging from one qualitative end point to another qualitative end point (e.g., strongly disagree to strongly agree). Likert items typically have 5 to 11 discrete response options.

Ordinal data do not have metric information. Although the response options might be numerically labeled as '1', '2', '3', …, the numerals only indicate order and do *not* indicate equal intervals between levels. For example, if the response items include '3' = "neither sad nor happy," '4' = "moderately happy," and '5' = "very happy," we cannot assume that the increment in happiness from '3' to '4' is the same as the

☆ The authors gratefully acknowledge constructive suggestions from Roger Giner-Sorolla, Wes Bonifay, and two anonymous reviewers.
* Corresponding author.
E-mail address: torrin.liddell@gmail.com (T.M. Liddell).

https://doi.org/10.1016/j.jesp.2018.08.009
Received 6 November 2017; Received in revised form 3 August 2018; Accepted 16 August 2018
Available online 07 September 2018

increment in happiness from ‘4’ to ‘5’.

Metric methods assume that the data are on an interval or ratio scale (Stevens, 1946, 1955). Interval scales define distances between points (not only ordering), and ratio scales furthermore specify a zero point so that ratios of magnitudes can be defined. We use the term *metric* to refer to either interval or ratio scales because the distinction between interval and ratio scale is immaterial for our applications. In metric data, the differences between scores are crucial. Thus, when metric models are applied to ordinal data, it is implicitly (and presumably incorrectly) assumed that there are equal intervals between the discrete response levels. As we will demonstrate, applying metric models to ordinal data can lead to misinterpretations of the data.

2.1. Ordinal data are routinely analyzed with metric models

We wanted to assess the extent to which contemporary researchers actually do use metric models to analyze ordinal data. By metric models, we mean models that assume a metric scale, including models underlying the *t* test, analysis of variance (ANOVA), Pearson correlation, and ordinary least-squares regression. We examined the 2016 volumes of the *Journal of Personality and Social Psychology* (JPSP), *Psychological Science* (PS), and the *Journal of Experimental Psychology: General* (JEP:G). All of these journals are highly ranked. Consider, for example, the SCImago Journal Rank (SJR), which “expresses the average number of weighted citations received in the selected year by the documents published in the selected journal in the three previous years, –i.e. weighted citations received in year X to documents published in the journal in years X-1, X-2 and X-3” (http://www.scimagojr.com/help.php, accessed May 15, 2017). In 2015, the most recent year available, the SJRs were 5.040 for JPSP (13th highest of 1063 journals in psychology, 3rd of 225 journals in social psychology), 4.375 for PS (18th highest in psychology, 8th of 221 journals in psychology-miscellaneous), and 3.660 for JEP:G (21st highest in psychology, 2nd of 118 journals in experimental and cognitive psychology).

We searched the journals for all articles that mentioned the word “Likert” anywhere in the article, using the journals' own web site search tools (http://journals.sagepub.com/search/advanced for PS, http://psycnet.apa.org/search/advanced for JPSP and JEP:G, all journals searched March 22, 2017). There may be many articles that use ordinal data without mentioning the term “Likert,” but searching for ordinal data using more generic terminology would be more arbitrary and difficult. The search returned 38 articles in JPSP, 20 in PS, and 20 in JEP:G, for a total of 78 articles. (A complete table of results is available online at https://osf.io/53ce9/.) Of the 78 articles, we excluded 10 because they did not actually use a Likert variable as a dependent variable (of the 10 articles excluded, 1 only referred to another article without using Likert data itself, 3 mis-used the term to refer to an interval measure, 2 used the term for scales with 100 or more response levels, 1 provided no analysis of the Likert data, and 3 used the Likert data only as a predictor and not as a predicted value). *Of the 68 articles, every one treated the ordinal data as metric and used a metric model; not a single analysis in the 68 articles used an ordinal model.*

Because it appears that the vast majority of applied researchers in the psychological sciences analyze ordinal data as if they were metric, we believe it is important to point out a variety of potential problems that can arise from that practice. We also illustrate analyses that treat ordinal data as ordinal, and that typically describe the data much more accurately than metric models.

2.2. Metric and ordinal models

To keep our examples and simulations straight forward, we use the most common versions of metric and ordinal models. When data are assumed to be on a metric scale, our models use a normal distribution

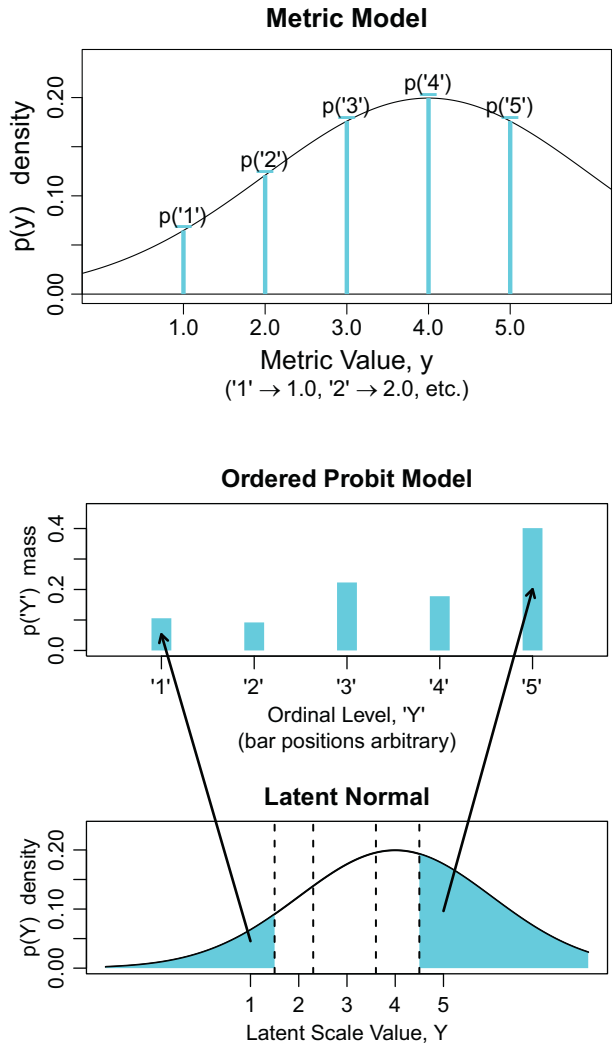

Fig. 1. *Upper panel: Metric model of ordinal data.* Ordinal values are mapped to corresponding metric values on the horizontal axis, with ‘1’→1.0, ‘2’→2.0, and so forth. The probability of each value is the normal density as shown by the heights of the lines. *Lower pair of panels: Ordered-probit model.* A latent scale on the horizontal axis is divided into intervals with thresholds marked by dashed lines. The cumulative normal probability in the intervals is the probability of the ordinal values, as suggested by the shading under the normal curve and arrows pointing to the corresponding probability in the bar plot.

for the residual noise. A normal distribution is assumed by the traditional *t* test, analysis of variance (ANOVA), linear regression, and so on. When data are instead assumed to be on an ordinal scale, our models use a thresholded cumulative normal distribution for the noise. A thresholded cumulative normal distribution is used by traditional “ordered-probit” models (e.g., Becker & Kennedy, 1992). The key difference between the metric-scale and ordinal-scale models is that the metric model describes a datum's probability as the normal probability density at a corresponding metric value, whereas the ordinal model describes a datum's probability as the cumulative normal probability between two thresholds on an underlying latent continuum.

Fig. 1 illustrates the difference between metric (normal density) and ordered-probit (thresholded cumulative normal) models. Suppose we have data from a Likert-response item, with possible ordinal values labeled ‘1’, ‘2’, ‘3’, ‘4’, and ‘5’. According to the metric model, shown in the upper panel of Fig. 1, the probability of ordinal response ‘1’ is the

normal probability density at the metric value 1.0, the probability of ordinal response '2' is the normal probability density at the metric value 2.0, and so on for the other levels. In this approach the analyst pretends that there are equal distances between the ordinal responses, and maps the ordinal responses to corresponding metric scale values. The upper panel of Fig. 1 shows the normal probability densities at these values.

On the other hand, according to the ordered-probit model shown in the lower pair of panels of Fig. 1, the ordinal responses are generated by chopping a normally distributed latent continuous value into subintervals. For example, suppose you are asked, "How happy are you? The response options are '1' = very unhappy, '2' = mildly unhappy, '3' = neutral, '4' = mildly happy, '5' = very happy." Intuitively, there is an underlying continuous scale for feeling of happiness, which is divided at some thresholds for mapping into discrete responses. The lower panel of Fig. 1 has the latent continuous value as its horizontal axis, and the normal distribution represents the variability of this latent value. The vertical dashed lines indicate the thresholds on the latent scale that divide ordinal response categories. The area under the normal curve between the thresholds is the probability of the corresponding ordinal response. The bar plot shows the cumulative-normal probabilities as bar heights. Arrows suggest the correspondences of areas under the normal curve to bar heights. When modeling the data, the analyst estimates the values of mean and standard-deviation parameters (μ and σ) in the normal distribution, and also the values of the thresholds. This sort of model is sometimes called an "ordered-probit," "ordinal probit," or "thresholded cumulative normal" model (e.g., McKelvey & Zavoina, 1975; Winship & Mare, 1984). This model is used as one of the standard models of ordinal data in both frequentist and Bayesian analysis (e.g., Albert & Chib, 1997; Kruschke, 2015; Lynch, 2007).

For both models in Fig. 1, the normal distributions are the same (i.e., they have the same mean and standard deviation) but the probabilities of the ordinal values are different. For instance, the probability of ordinal value '5' is relatively large in the ordered-probit model because its probability is given by the cumulative probability under the entire normal distribution above the highest threshold. But in the metric model the probability of ordinal value '5' is given by the normal density at 5.0. As another comparison, consider the relative probabilities of ordinal values '3' and '4'. In the metric model, p('3') is less than p('4'), but in the ordered-probit model p('3') is greater than p('4') because of the cumulative areas between thresholds that happen to be unequally spaced in this example.

Now we define a few key terms that are needed to clearly describe the analyses. We defer complete mathematical details to the Appendix. The metric model uses a normal distribution that has a mean parameter, μ (mu), and a standard deviation parameter, σ (sigma). There is a distinct mean and standard deviation for each group. The ordered-probit model also uses a normal distribution on a latent scale, so the ordered-probit model also estimates a distinct mean and standard deviation for each group.

The only additional parameters in the ordered-probit model are the threshold values between intervals, denoted θ_k. The thresholds are assumed to be controlled by the anchor labels of the response prompt (e.g., "very happy," "mildly happy," and so on), which are the same for all groups. Therefore the same thresholds are used by all groups. Because of algebraic trade-offs, the outer thresholds of the ordered-probit model are fixed and only the interior threshold are estimated from the data. For instance, for data with K ordinal levels, both the metric model and ordered-probit model estimate a mean (μ) and standard deviation (σ) for each group, while the ordered-probit model estimates only $K - 3$ additional thresholds ($\theta_2,\ldots,\theta_{K-2}$).

In our analyses, we use Bayesian methods to estimate the parameter values in both the metric models and the ordered-probit models. We use Bayesian methods because of their flexibility to specify models with the properties we want (such as unequal variances across groups, and hierarchical structure on variances) and because of the accuracy and richness of the information provided. Bayesian analysis provides a probability distribution over the parameter space, so that we can see what parameter values are most probable, given the data. This probability distribution over parameters is called the posterior distribution because it is derived "after" considering the data. The spread of the posterior distribution indicates the uncertainty of the estimated value, and this uncertainty is expressed by the *highest density interval (HDI)*. The 95% HDI includes 95% of the parameter distribution, such that all values inside the HDI have higher probability density that values outside the HDI. In other words, the 95% HDI spans the 95% most credible values of the parameter. For each parameter value we estimate, we will report its posterior modal value and 95% HDI. These values are roughly analogous to the maximum likelihood estimate and 95% confidence interval in frequentist statistics. A non-technical introduction to Bayesian analysis is provided by Kruschke and Liddell (2018a) and a comparison with frequentist approaches is presented by Kruschke and Liddell (2018b). Complete details of the models are provided in the Appendix.

2.3. Is treating ordinal data as metric innocuous?

Aside from not having equal distances between levels, ordinal data also routinely violate the distributional assumptions of metric models. Traditional metric models such as t tests assume normally distributed data (around the predicted central tendencies). But real ordinal data (if assumed to fall on a metric scale) are often strongly skewed, heavy-tailed or thin-tailed, or multi-modal. Thus, treating ordinal data as if they were normally-distributed equally-separated metric values is not appropriate. But does the practice actually lead to problems? Or, is the practice innocuous and desirable for its simplicity?

We demonstrate that the practice can lead to systemic errors, hence it is not innocuous. Ordinal models yield better descriptions of the data, and therefore analysts should prefer ordinal models for ordinal data over metric models. In this article we show that describing ordinal data with a metric model can lead to serious misinterpretations of the data. Our simulations generate ordinal data from ordered-probit models and then analyze the data using both metric and ordered-probit models. We show that the ordered-probit models accurately recover the true generating parameters, while the metric models systematically yield erroneous interpretations. We also show several cases of real-world data with characteristics very similar to the simulated data, to demonstrate that these situations arise in real research data. Most importantly, we explain why there are discrepancies between metric and ordered-probit models in principle, and illustrate how an infinity of configurations create false alarms (Type I errors), misses (Type II errors), and inversions of effects.

3. Examples of errors when treating ordinal data as metric

The simple examples of this section introduce a few of the erroneous inferences that can be made when treating ordinal data as if they were metric. To be able to declare an inference to be an error, we have to know the correct answer. Therefore in this section we generate simulated data from ordered-probit models that have known parameter values. We show that analyses with ordered-probit models recover the true generating parameter values well (subject to random sampling variation), whereas analyses that treat the data as if they were metric yield inaccurate estimates. In particular, for metric models we illustrate a false alarm (i.e., a Type I error) when the true parameter values are set to their null values, and we illustrate a failure to detect a non-zero effect (i.e., a "miss" or Type II error). Later we will show inversions of estimated means, in which the means estimated by the metric model are

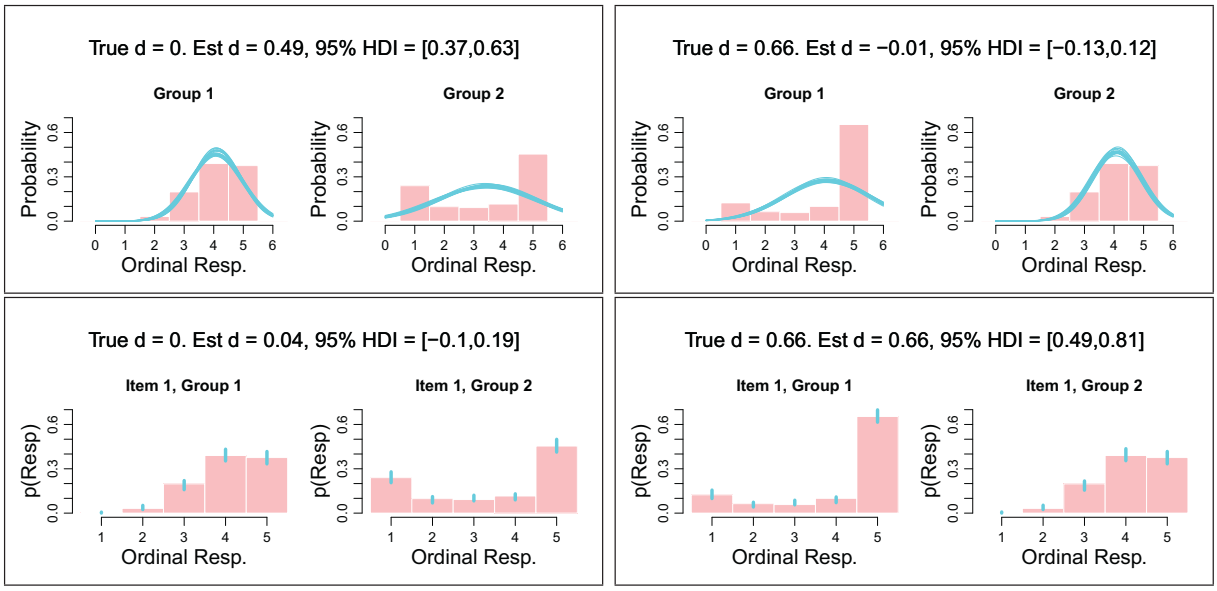

Fig. 2. Ordinal data from a single item for two groups, displayed in histograms (same data in upper and lower panels). The left column shows a false alarm (Type I error) by the metric model (corresponding to points Ⓐ and Ⓑ in Fig. 4). The right column shows a miss (Type II error) by the metric model (corresponding to points Ⓑ and Ⓑ in Fig. 4). Posterior predicted data probabilities are shown as in the upper panels has a smattering of normal curves and in the lower panels as dots. The metric model (normal curves) are a poor description of the data, but the ordered-probit model accurately describes the data probabilities.

in reversed order relative to the true generating means.

To generate simulated data from the ordered-probit model, we randomly sampled from normal distributions using $N = 500$ in each group, as an arbitrary but reasonably large sample size to recover the generating parameter values. The *effect size* between groups is measured as Cohen's d (Cohen, 1988), defined as the standardized difference between means: $d = (\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$. For simplicity, and to avoid concerns that the results depend on particular spacing of the thresholds, the thresholds were evenly spaced, with $\theta_1 = 1.5$, $\theta_2 = 2.5$, $\theta_3 = 3.5$, and $\theta_4 = 4.5$. Further details of the models are provided in the Appendix.

The left side of Fig. 2 shows a case of equal underlying means in the true ordered-probit model, that is, a true effect size of zero. The ordered-probit model accurately recovers the generating parameters. But the metric model badly mis-estimates the effect size. In other words, this is a gross Type I error or false alarm by the metric model. Moreover, the ordered-probit model accurately describes the distribution of ordinal responses within each group, as shown by dots superimposed on the data histograms. By contrast, the metric model poorly describes the distribution of ordinal responses within each group, as shown by the (poorly fitting) normal distributions superimposed on the data histograms.

The right side of Fig. 2 shows a case of *un*equal underlying means in the true ordered-probit model, that is, a case in which the true effect size is large (non-zero). The ordered-probit model accurately recovers the generating parameter values. But the metric model estimates the effect size to be nearly zero. In other words, this is a gross Type II error or "miss." Again, the ordered-probit model accurately describes the distribution of ordinal responses within each group, while the metric model (normal distribution) is a poor description.

We show in the next section that it is also trivial to generate an infinite number of cases in which the two models show very different means *in opposite directions*. We will subsequently show many examples of real data with these dramatic inconsistencies between metric and ordered-probit models.

4. Why the metric model fails

In this section we present the key concepts of the article that explain the mismatch between the ordered-probit and metric models. The explanation centers on understanding the relationship between the latent mean in the ordered-probit model to the mean of ordinal values in the metric model.

Fig. 3 shows the mean of the ordinal values as a function of the latent mean (mu), for particular values of latent standard deviation (SD, sigma) and thresholds (theta's). The figure assumes an ordinal-value range of 1 to 5 merely for illustration, and the exact ordinal range makes no qualitative difference to the phenomena reported below. For any particular values of the latent parameters, the mean of the ordinal values was computed as the mathematically exact expected value (see Eq. (4) in the Appendix). Notice that the mean of ordinal values is an S-shaped (sigmoidal) function of the latent mean, with the function reaching asymptotes at the low and high ordinal values. This squeezing of the latent mean into a limited range makes intuitive sense because the highest ordinal value is 5 no matter how large the latent mean is, and the lowest ordinal value is 1 no matter how low the latent mean is. Thus, the ordinal values censor extreme latent values, and the terminal ordinal values lose information about how extreme the latent value is.

Fig. 4 shows curves for two choices of the latent standard deviation (SD. σ). The curve for the smaller latent SD is steeper than the curve for the larger SD. This makes intuitive sense if one considers a latent mean at, say, 5: A small SD implies that only the ordinal values near 5 will have high probability and the mean of the ordinal values will be near 5, whereas a large SD implies that a broad range of ordinal values will have notable probability and the mean will be farther from (i.e., lesser than) 5.

Consider in Fig. 4 the parameter combinations indicated by the letter-labeled points. Each letter-labeled point represents a particular combination of μ and σ and corresponds to a particular pattern of ordinal data that would be generated by those parameter values in the ordered-probit model. The points indicated by Ⓐ and Ⓑ represent two groups with the *same* latent mean. Notice that the means of the

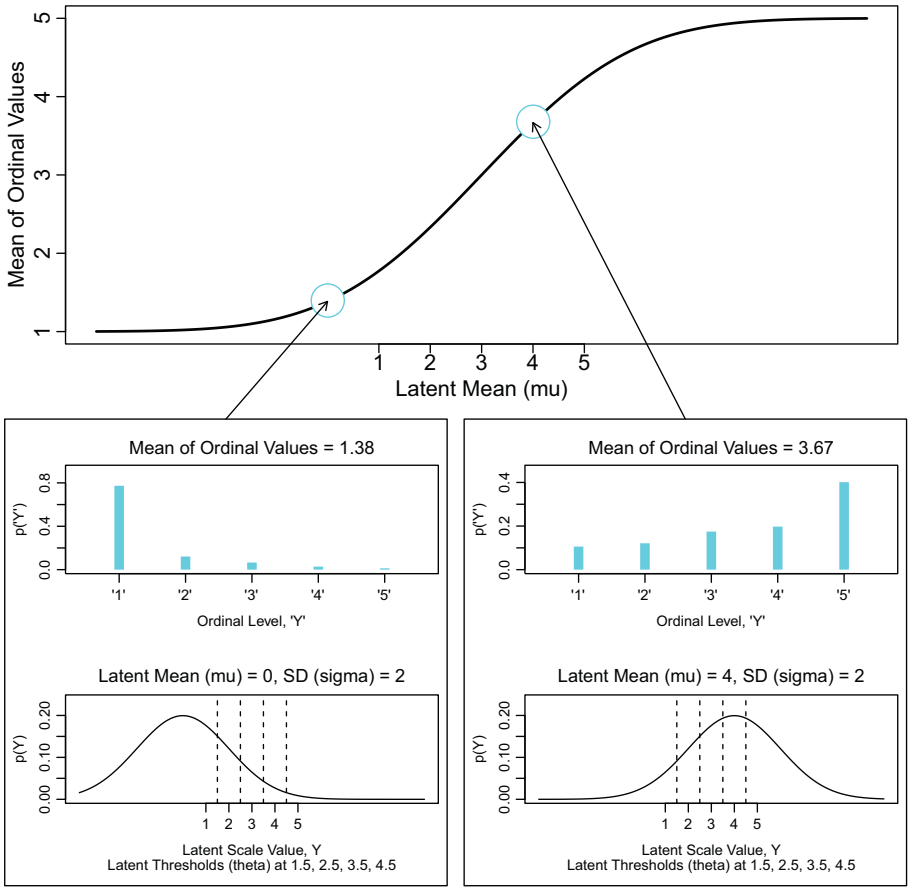

Fig. 3. *Upper panel:* The mean of the ordinal values is plotted as a function of the latent mean (mu). *Lower panels:* For the two marked points on the sigmoidal curve, lower boxes show the corresponding latent normal distribution and resulting ordinal values. The only difference between the two lowest panels is the mean (mu) of the latent normal distribution; the two lowest panels have the same latent standard deviation (sigma) and the same latent thresholds (theta's).

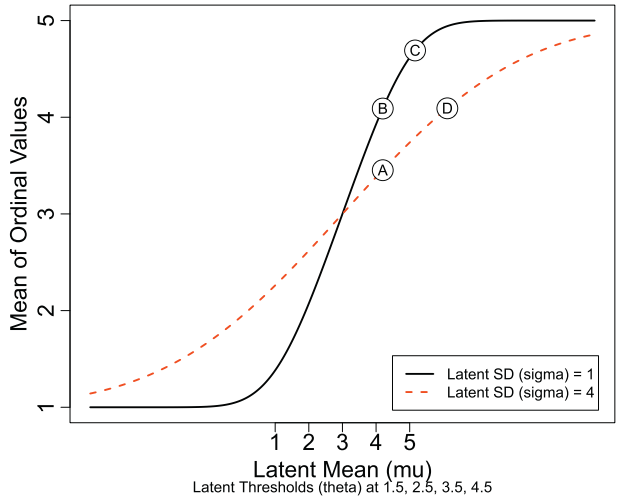

Fig. 4. Ordinal mean as a function of latent mean (mu) and SD (sigma). Groups marked Ⓐ and Ⓑ illustrate a false alarm (Type I error) for which the underlying means are exactly equal but the ordinal means are very different. Groups marked Ⓑ and Ⓓ illustrate a miss (Type II error) for which the underlying means are quite different but the ordinal means are exactly equal. Groups marked Ⓒ and Ⓓ illustrate an inversion for which the underlying means have $\mu_D > \mu_C$ but the ordinal means incorrectly have $\mu_C > \mu_D$.

resulting ordinal values are *not* equal. The metric model estimates the means of the ordinal values, and therefore produces a false alarm, or Type I error. This case was exemplified in Fig. 2.

Conversely, points Ⓑ and Ⓓ in Fig. 4 represent two groups with *equal* ordinal means, but *un*equal latent means. Because the metric model reveals the means of the ordinal values, the metric model will miss the true non-zero latent difference, that is, the metric model commits a Type II error. This case was exemplified in the right side of Fig. 2.

Consider what happens if two groups happen to have identical latent variances. This situation is illustrated in Fig. 4 by any two points on the same sigmoid, such as Ⓐ and Ⓓ. The true separation of latent means on the horizontal axis is much larger than the sigmoidally-compressed separation of ordinal means on the vertical axis (notice the numerical scales on the axes). This compression of differences tends to reduce statistical power for finite sample sizes. We illustrate this situation with real data later in the article.

Finally, consider groups indicated by points Ⓒ and Ⓓ in Fig. 4. In this case, the direction of difference indicated by the metric model is *opposite* the true direction of difference in the latent means. Later we present real data with many cases of inversions between ordinal and latent means (along with many false alarms and misses).

The same phenomena occur for *un*equal bin widths between thresholds. The only change with unequal bin widths is a change in the exact shape of the sigmoids and consequently the exact magnitude of the effects at particular underlying scale values. The qualitative problems are the same regardless of interval spacing.

In summary, multiple types of problems can occur when treating ordinal data as if they were metric. These problems include increased false-alarm rates (as illustrated by Ⓐ vs Ⓑ in Fig. 4), low correct-detection rates (as illustrated by Ⓑ vs Ⓓ in Fig. 4), and distorted or even reversed effect-size estimates (as illustrated by Ⓒ vs Ⓓ in Fig. 4).

Fig. 4 captures the core concepts of this article. It is crucial to understand that Fig. 4 implies there are an infinite number of

combinations of underlying parameter values that generate false alarms, or misses, or inversions. *Any* two distinct variances create two distinct sigmoid curves as in Fig. 4. A vertical slice through the two curves at *any* point along the horizontal axis (except at the infinitesimal point where the sigmoids intersect) creates combinations analogous to Ⓐ and Ⓑ that yield false alarms (Type I errors). A horizontal slice through the two curves at *any* point along the vertical axis (except at the infinitesimal point where the sigmoids intersect) creates combinations analogous to Ⓑ and Ⓓ that yield misses (Type II errors). A downward-right diagonal slice through the two curves at *any* location (except at the infinitesimal point where the sigmoids intersect) creates combinations analogous to Ⓒ and Ⓓ that yield inversions of apparent and true means. These misrepresentations are inherent in the mathematics and cannot be avoided. Importantly, there is no need to run Monte Carlo simulations to illustrate the errors because estimated parameter values converge to the true parameter values shown in Fig. 4 as sample size gets large. Examples of this were presented in Fig. 2, which used samples with $N = 500$.

5. Movie ratings: real data instantiate these problems

A rich source of ordinal data is star ratings in reviews of products or services posted on the internet. We went to Amazon.com and in the "Movies & TV" category entered the search term, "drama." The results were sorted by "relevance" as defined by Amazon.com. We then recorded the star ratings of the first 36 movies that had at least 500 reviews; the number of reviews ranged from 502 to 111,232. We chose 36 movies merely because it is a large enough set to illustrate a variety of phenomena but small enough to display all data in one figure. We chose movies merely because there are many movies with many reviews; however, the same phenomena occur for other products such as cell phones, etc. At Amazon.com, the ratings summary of a movie consists of the number of reviews and the percentage given to each star rating from 1 to 5 stars. The percentages were provided only to whole-digit precision. To compute the frequency of each star rating, we multiplied the total number of reviews by the percentages and rounded to the nearest whole number.

We analyzed the data using an ordered-probit model and using a corresponding metric model. The ordered-probit model estimated thresholds for cutting the latent normal distribution, and used the bin probabilities for the describing the ratings. The metric model was identical to the ordered-probit model except the metric model used a normal density to describe the ratings (with '1'→ 1.0, '2'→ 2.0, and so forth). The metric model had 72 parameters (36 means plus 36 standard deviations) while the ordered-probit model had 74 parameters (all of the metric-model parameters plus 2 thresholds). Complete details are in the Appendix.

Fig. 5 shows data from the movies fit by the ordered-probit model. Fig. 6 shows the same data fit by the metric model. By comparing the figures, it is visually obvious that the ordered-probit model fits the data *much* better than the metric model.

Fig. 7 shows the means of the metric model plotted against the means of the ordered-probit model. Importantly, notice the non-monotonicities in the plot: The means are ranked differently by the two models. For these real data, we cannot assert that the ordered-probit model is the "true" model, and therefore we cannot assert that a discrepancy from the ordered-probit model is an error. But clearly the ordered-probit model is a much better description of the data than the metric model, so we can point out configurations of means that are analogous to false alarms like Ⓐ and Ⓑ in Fig. 4, in which the means in the ordered-probit model are nearly equal while the means in the metric model are quite different. Many pairs of cases show this sort of false alarm (Type I error), including 20–34, 3–11, 3–4, 6–29, 12–16, etc.

There are also many pairs of means that resemble the misses (Type II errors) of Ⓑ and Ⓓ in Fig. 4, in which the latent means of the ordered-probit model are quite different, but the means of the metric model are nearly equal. Pairs that show this sort of error include 2–10, 1–34, 20–29, 2–24, 12–19, 8–28, etc.

There are also many pairs of means that resemble the inversions of Ⓒ and Ⓓ in Fig. 4, in which the means in the ordered-probit model are quite different, as are the means in the metric model, but *in opposite directions*. Pairs that show this inversion include 6–13, 5–35, 20–35, 5–14, 3–34, 4–34, 2–12, 2–16, etc.

In particular, Fig. 8 shows an example of two movies (Cases 5 and 6) that are ranked differently by the two models. The two models show differences between the two movies in opposite directions, in both models quite strongly. The ordered-probit model is a much better description of the data than the metric model, and therefore we should treat the ordered-probit parameter values as more meaningful.

A simple script for this analysis, in the R computer language, is provided at https://osf.io/53ce9/. The script allows the user to read in any similarly structured data file (i.e., one item measured in multiple groups) and produces detailed results including graphs like Figs. 5 through 8.

[Blank space intentionally inserted for formatting of subsequent material.]

6. Equal metric SDs do not avoid problems

Based on the examples highlighted in Fig. 4, one might surmise that the problems can be avoided if the variances of the *ordinal* values of the groups happen to be equal. Unfortunately, the variances of the ordinal values do not reveal the underlying relation of variances of the latent values. A thorough explanation of the general situation is provided in the Appendix.

Despite the lack of correspondence between metric-model SD and ordered-probit-model SD, it is true that *if* the theoretical metric-model σ's (of ordinal values generated from the ordered-probit model) are equal, then the metric-model means are monotonically related to the underlying ordered-probit means, but the metric-model effect size is less than the ordered-probit model effect size. (A detailed example is provided by Fig. 14 in the Appendix.) Because real data have finite samples, the ordered-probit model is more powerful at detecting differences of means than the metric model when the metric σ's are equal. Examples of this improved sensitivity by the ordered-probit model are evident by comparing movie cases 2 vs 25, cases 11 vs 34, and cases 29 vs 35, etc. Fig. 9 shows the comparison of movies 35 and 29. The metric σ is nearly 1.2 for both movies. Both models show that the mean of movie 35 is greater than the mean of movie 29, but the ordered-probit model clearly indicates that the difference is non-zero (because the 95% HDI falls far away from a difference of zero) while the metric model includes a difference of zero within its 95% HDI. The ordered-probit model is more powerful and more accurately describes the data.

7. Average of multiple ordinal items

Consider a situation in which respondents provide ratings on four similar items. For example, a questionnaire might ask the following four questions: How happy are you?, How sad are you? (reverse scaled), How satisfied are you?, and How disappointed are you? (reverse scaled). Responses to the four items tend to be very highly correlated, that is, if a respondent rates one item low the respondent will tend to rate all items low, and if a respondent rates one item high the respondent will tend to rate all items high. When intercorrelation of the items is high, and especially when the items are meaningfully related, analysts routinely take the average rating of the items. Notice that taking an arithmetic average is already making the assumption that the ordinal values can be treated as metric. The averaged values are then

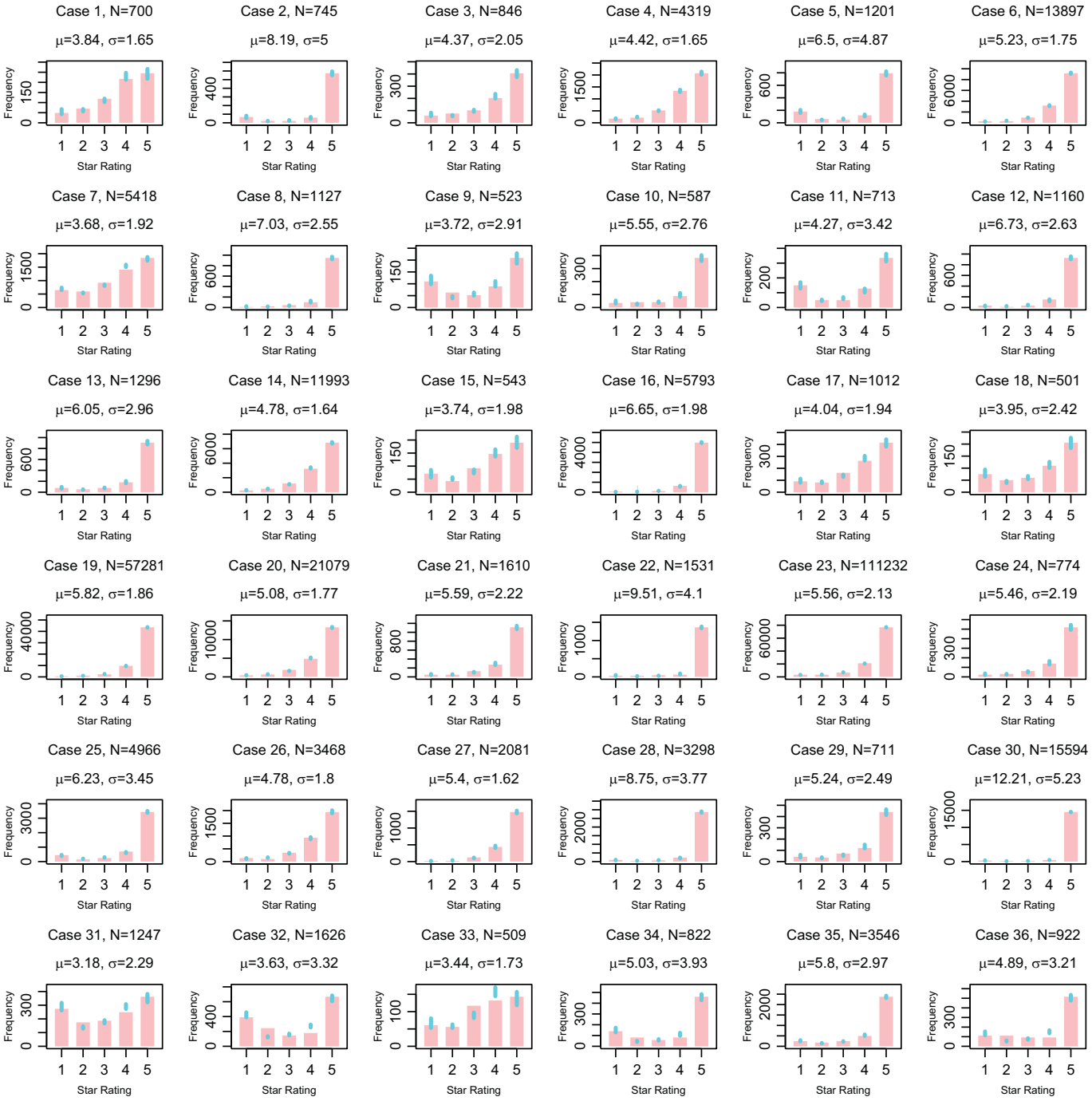

Fig. 5. Ratings data from 36 movies (indicated by the Case number) are shown as histograms. Posterior predicted probabilities from the ordered-probit model are superimposed on the data as dots on each histogram bar. Each dot also has a small vertical segment that indicates the 95% HDI of the predicted probability, but the segments are sometimes smaller than the dot and are therefore not visible. The fit of the ordered-probit model is remarkably good, as the predictions match the data for a variety frequency distributions across cases.

put into standard metric analyses (such as the *t* test).

Some authors have argued that, despite the ordinal character of individual Likert items, averaged ordinal items can have an emergent property of an interval scale and so it is appropriate to apply metric methods to the averaged values (e.g., Carifio & Perla, 2007, 2008). It is intuitively plausible that the averaging could produce data that at least *look* more continuous and therefore may not suffer from the problems pointed out above. Unfortunately that intuition is wrong. We show in this section that an average of ordinal items has the same problems as a single item.

7.1. An ordered-probit model for correlated ordinal items

Suppose there are several related Likert items. We will assume that there is a single latent variable that underlies all the items simultaneously. (A more elaborate model would use a distinct latent variable

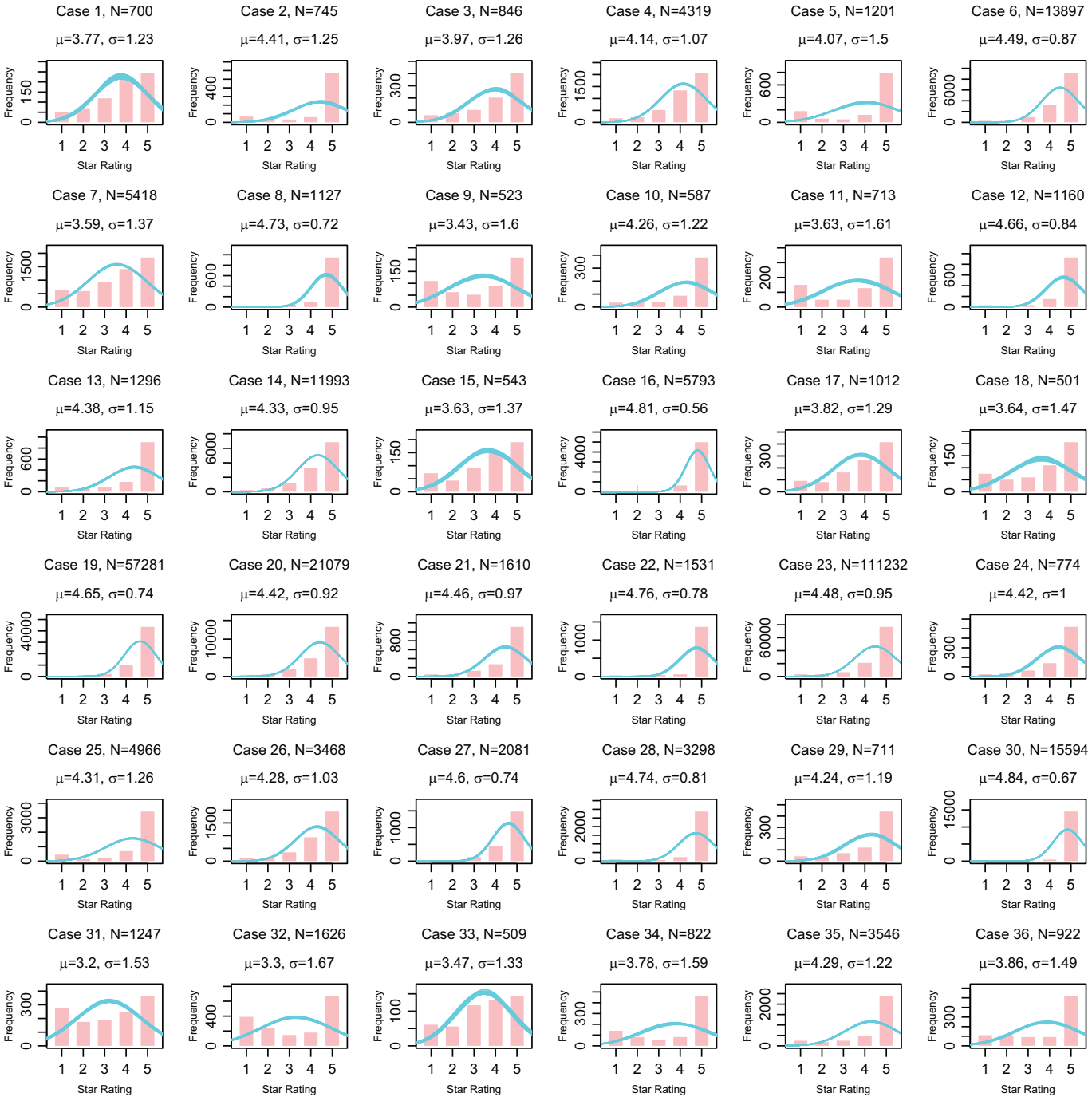

Fig. 6. Ratings data from 36 movies (indicated by the Case number) are shown as histograms. Posterior predicted normal distributions from the metric model are superimposed on the data. The normal distributions are a smattering of representative parameter combinations from the posterior distribution, so the line thickness of the normal distributions indicates the uncertainty of parameter estimate. The fit of the metric model is poor, as the data histogram bars protrude above and below the normal distributions quite severely.

for every item, with the latent dimensions strongly correlated across responders, but this model is difficult to implement for technical reasons.) As before, each group is assumed to be normally distributed on the latent variable but the latent value is mapped to each item's ordinal output by different item-specific thresholds. For details, see Eq. (3) in the Appendix. As before, because of trade-offs in parameter values that yield the same response probabilities, we fix the lowest and highest thresholds of the first item. Notice that the data being modeled by the ordered-probit model are the ratings of all the individual items, whereas the data being modeled by the metric model are the averaged ordinal values.

7.2. Examples of errors with averaged ordinal items

To generate simulated data for correlated items, we sampled from a multivariate normal distribution that had correlations of 0.8 for all

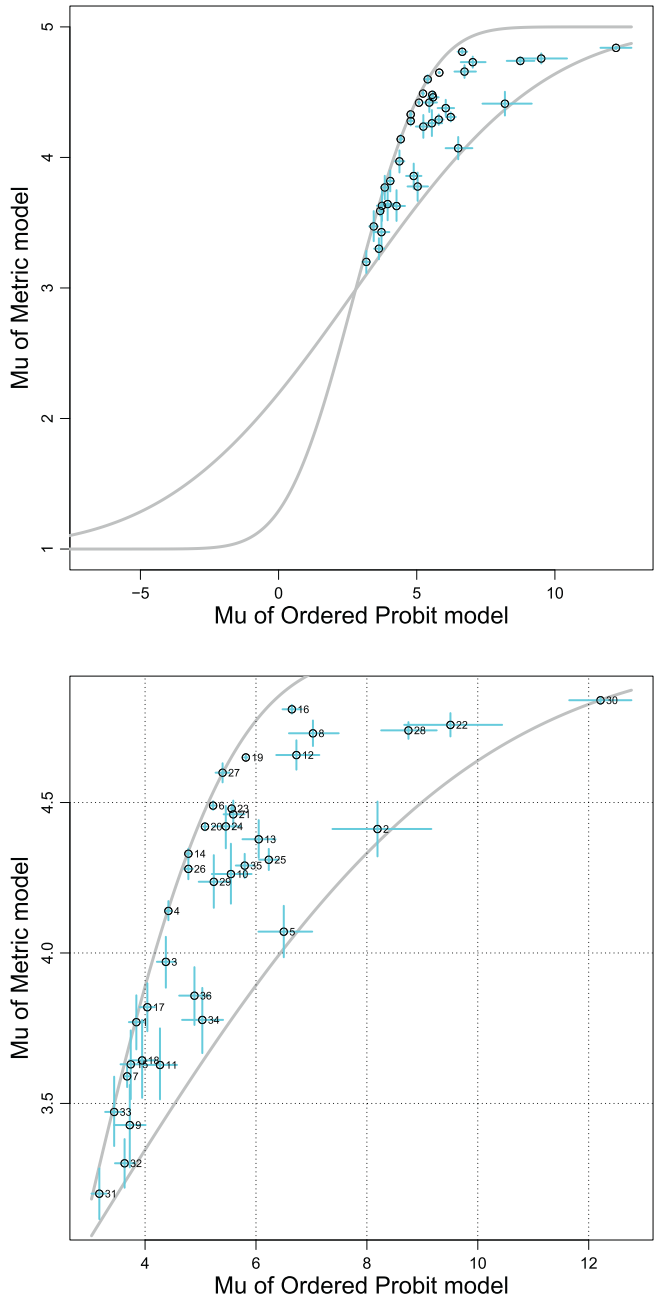

Fig. 7. Movie data: Posterior modal mu values from the metric model are plotted against the posterior modal mu values for the ordered-probit model. Upper panel shows full axis range analogous to Fig. 4. Lower panel zooms in with each dot labeled by its case number. Segments intersecting the dots indicate the 95% HDIs of the parameters. Curves are the S-shaped plots of metric mu as a function of ordered-probit mu for the smallest latent sigma (upper curve) and largest latent sigma (lower curve).

pairwise combinations of dimensions. Each dimension corresponds to an item. For simplicity, and to avoid concerns that the results depend on particular spacing of the thresholds, the thresholds for every dimension were set equal to each other and evenly spaced. The means and standard deviations were the same on every dimension (item), but each group had its own mean and standard deviation just as in the previous examples. We used $N = 500$ in each group, as an arbitrary but reasonably large sample size to recover the generating parameter values. When analyzing the data with an ordered-probit model, we used the model described previously. In other words, the model assumed a single latent dimension instead of multiple correlated latent dimensions.

We first illustrate results for two groups that have a true effect size of zero, $d = 0.0$, corresponding to points Ⓐ and Ⓑ in Fig. 4. Recall that Fig. 2 illustrated this case for a single ordinal item. Fig. 10 shows results from three ordinal items. The upper panels of Fig. 10 show histograms of the averaged ordinal responses. Notice that the averaged ordinal items can take on only a discrete set of values. For example, if there are Q items and each item has responses $1,\ldots,K$ then the average can only have values Q/Q, $(Q + 1)/Q$, $(Q + 2)/Q$, …, KQ/Q. In our example there are three items each with five ordinal levels, so average responses can have discrete values of 3/3, 4/3, 5/3, 6/3, and so on. Notice in Fig. 10 that the metric model of the averaged ordinal values greatly overestimates the effect size, while the ordered-probit model accurately recovers the true effect size of zero. Moreover, the metric model is a poor description of the averaged ordinal values, while the ordered-probit model accurately describes the probabilities of every ordinal value in every item.

Thus, averaging across items does not solve the problem of false alarms created when analyzing ordinal data with a metric model. Moreover, when the number of items being averaged increases, the problem is not ameliorated. The erroneous magnitude of the effect size remains the same with more items, as can be seen by comparing the metric model results for a single item in Fig. 2 with three items in Fig. 10.

Next, we illustrate results for two groups that have a true effect size greater than zero, corresponding to points Ⓑ and Ⓓ in Fig. 4. The right column of Fig. 10 shows results from three ordinal items. The upper-right panel of Fig. 10 show histograms of the averaged ordinal responses, where it can be seen that the metric model of the averaged ordinal values estimates the effect size to be nearly zero. By contrast, the ordered-probit model accurately recovers the true effect size. Moreover, the metric model is a poor description of the averaged ordinal values, while the ordered-probit model accurately describes the probabilities of every ordinal value in every item. Again we see that averaging across items does not solve the problem of missed effects (Type II errors) created when analyzing ordinal data with a metric model.

Moreover, when the number of items being averaged increases, the problem is not ameliorated. The erroneous magnitude of the effect size remains the same with more items, as can be seen by comparing the metric model results for a single item in Fig. 2 with three items in Fig. 10. Examples with six items show the same effects (not displayed here to conserve space). In general, more items, like larger sample sizes, will home in on the underlying generating values, such as the letter-labeled points in Fig. 4. The use of more items does not remove the core distortions in Fig. 4.

8. Discussion

8.1. Interactions and trend analysis

The problems with using metric models for ordinal data do not end with the problems pointed out above. The same underlying distortions will produce analogous problems in more complex designs and analyses.

Consider, for instance, four groups in a two-by-two factorial design. If two of the show inverted means, like groups Ⓒ and Ⓓ in Fig. 4, and two other groups are inverted in the opposite direction, then the metric model and the ordered-probit model will show cross-over interactions

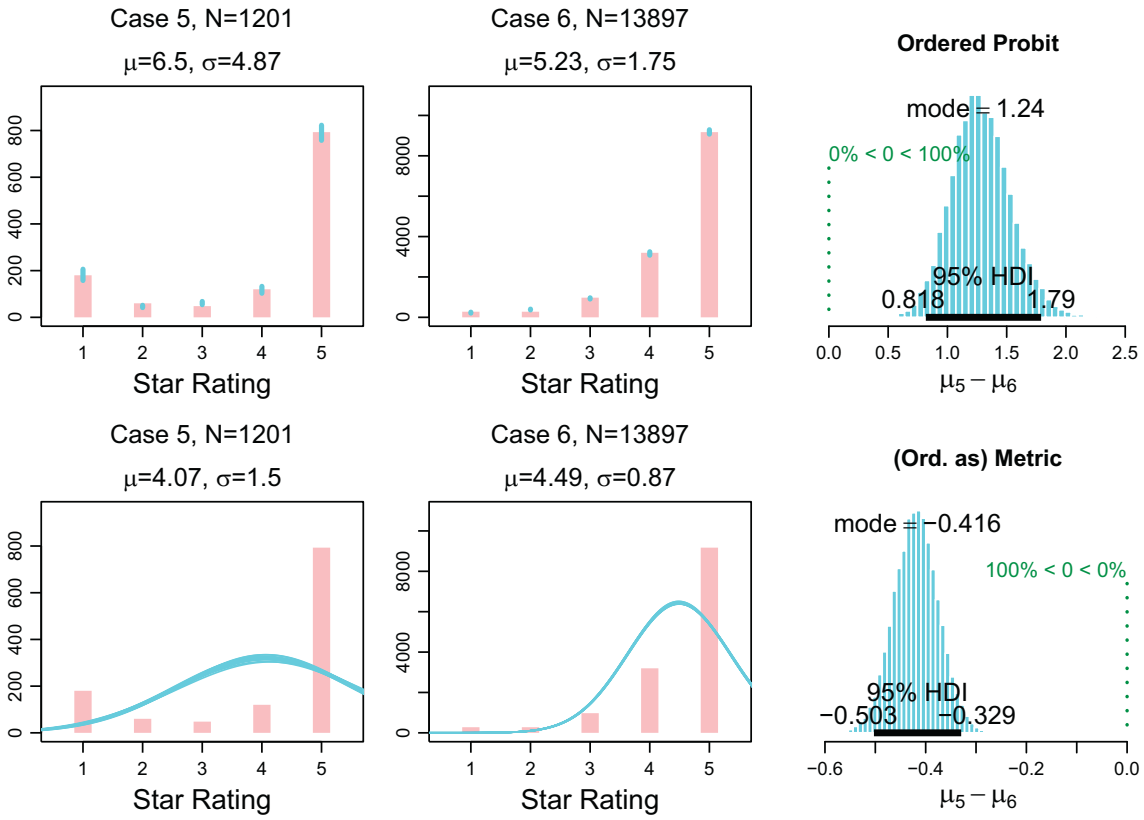

Fig. 8. Posterior difference of mu's for two movies (Cases 5 and 6). Upper row shows difference in ordered-probit model; lower row shows difference in metric model. The differences are strongly *in opposite directions*. The right panels shows the posterior distribution of the difference of means. Each posterior distribution is marked with a dotted line at a difference of zero, and the line is annotated with the percentage of the distribution below zero and above zero. Notice the ordered-probit model fits the data much better.

in opposite directions. Examples and figures are presented in the Appendix.

Trend analyses can also be badly distorted because of the sigmoidal squashing of underlying trends. A linear underlying trend in the ordered-probit model can appear to be non-linear in a metric model, or vice versa. Two linear trends in the ordered-probit model that have equal slopes can appear to have different slopes in the metric model. Examples and figures are presented in the Appendix.

8.2. Argument by denial: driving drunk

We have shown that there are many problems endemic in applying metric models to ordinal data. Nevertheless, some readers may dismiss these problems as possible in principle but so rarely encountered in practice that it is not worth the effort to move away from comfortable and familiar metric models. Unfortunately, we can not know in advance if any given set of data will yield different conclusions when analyzed with a metric model and an ordered-probit model. Moreover, we *do* know in advance that an ordered-probit model will better describe the data than a metric model. Arguing that it is okay to use a metric model instead of an ordered-probit model because a difference in conclusions may be rare is like arguing it's okay to drive drunk because accidents rarely happen. We are not suggesting that ordered-probit models are the *correct* model of ordinal data. We are arguing that ordered-probit models are clearly better than metric models of ordinal data, and that metric models *in principle* mis-estimate ordered-probit parameters regardless of the magnitude of their mis-estimation.

Some readers may wish for a test that would decide whether the inappropriateness of a metric model is small enough that they can continue to use it, much like a blood-alcohol test that would legally permit them to drive despite having a few drinks. As far as we can determine, such a statistical test would inherently involve fitting an ordered-probit model (or similar). And if an ordered-probit model is already implemented, it should be used instead of the metric model. To ease the transition to ordered-probit models, we have provided R scripts at https://osf.io/53ce9/ for analyzing grouped data (structured like the movie-rating example of Fig. 5).

8.3. Previous research did not flag these problems

A variety of previous investigators have examined false alarm rates in metric analyses of ordinal data (e.g., Boneau, 1960; Glass, Peckham, & Sanders, 1972; Havlicek & Peterson, 1976; Heeren & D'Agostino, 1987; Hsu & Feldt, 1969; Norman, 2010; Pearson, 1931). In general, they found false alarm rates not to be badly inflated. However, this body of work did not investigate circumstances we have highlighted that do produce false alarms. For example, cases with unequal variability across groups and means closer to one or the other end of the ordinal scale (such as Ⓐ and Ⓑ in Fig. 4) were not investigated in these papers, but we have demonstrated such cases do produce inflated false alarm rates.

Some previous researchers have pointed out problems when analyzing ordinal data with metric models. Nanna and Sawilowsky (1998) compared the performance of the *t* test and a corresponding

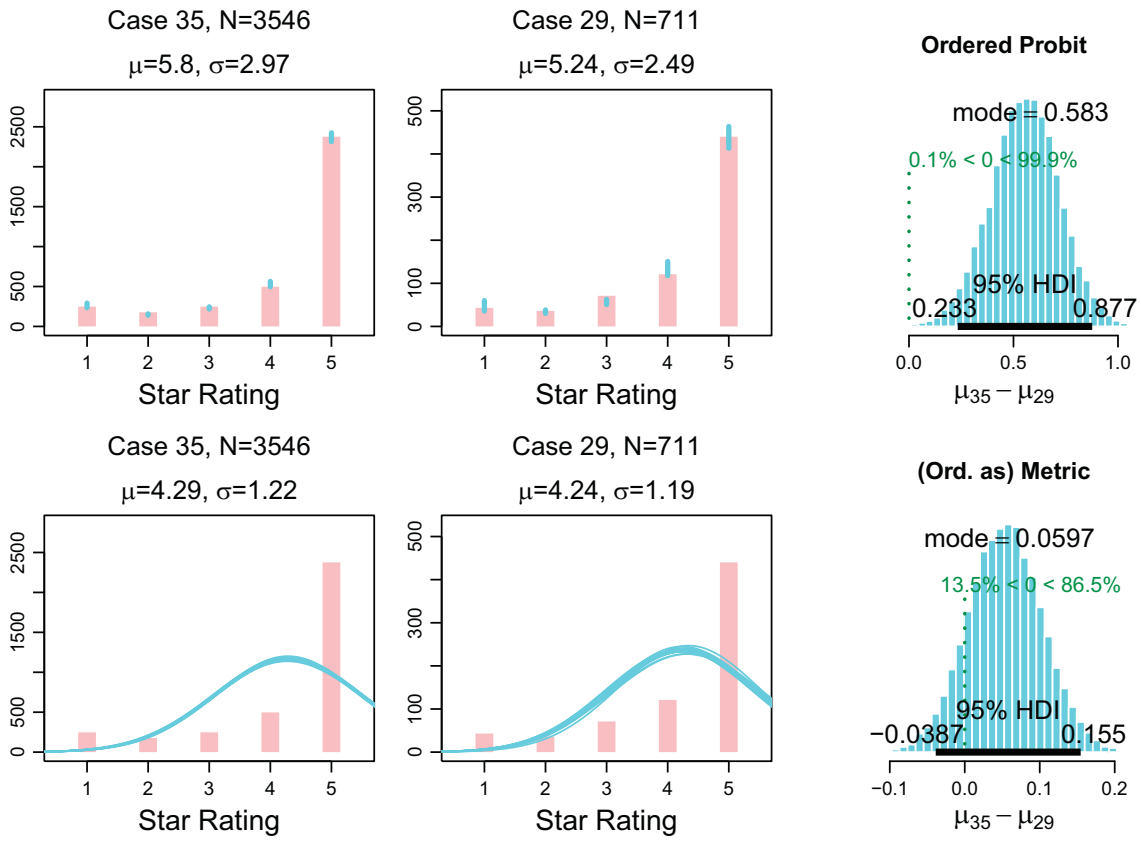

Fig. 9. An example in which the metric model shows cases of nearly identical SDs (both nearly 1.2). Although both models show that the mean of movie 35 is greater than the mean of movie 29, the ordered-probit model clearly indicates that the difference is non-zero (because the credible interval falls far away from a difference of zero) while the metric model includes a difference of zero within its credible interval.

nonparametric test (the Wilcoxon rank-sum test) on real-world data. The results indicated that regardless of sample size, the *t* test had a lower correct detection rate than the nonparametric test. This was true for both single Likert items and averaged multiple items.

Analytical work by Pearson (1931) suggested that the ANOVA can have a greatly reduced correct detection rate when the normality assumption is violated. To the extent that evidence regarding the effect of non-normality is generalizable to the specific case of ordinal data, these studies imply that the correct detection rate can be decreased when applying metric methods to ordinal data.

Additional literature has suggested that the application of metric methods to ordinal data produces distortions in estimates of effect size and uncertainty. In the context of Pearson's correlation, O'Brien (1979) investigated distortions in the estimated correlation due to ordinal data. Unlike the work of Havlicek and Peterson (1976) mentioned above, O'Brien was concerned with the actual magnitude of the estimated correlation, not merely the false alarm rate. O'Brien (1979) demonstrated via Monte Carlo simulation that Pearson's *r* can be substantially distorted when applied to ordinal data, at least when the distribution is skewed.

There has also been a large amount of work investigating the accuracy of factor analysis methods when applied to ordinal data, treating it as though it is metric. This work largely concludes that this practice leads to mis-estimation of factor correlations as demonstrated by both mathematical analysis (Mooijaart, 1983) and Monte Carlo simulation (Babakus, Ferguson, & Jöreskog, 1987; Dolan, 1994; Flora & Curran, 2004; Lubke & Muthén, 2004; Muthén & Kaplan, 1985). However, some authors (Flora & Curran, 2004; Muthén & Kaplan, 1985) provide evidence that, except for highly skewed ordinal data, the expected distortion is small.

In summary, a number of previous articles have argued that metric models of ordinal data do not badly inflate false alarm rates, but those researchers did not explore the range of cases we have systematically explained in Fig. 4. There is also a literature that suggests applying metric methods to ordinal data distorts effect sizes across various types of statistics and comparison methods, at least in some cases. We believe the analysis and graphical explanation we have presented in this article unifies these disparate cases and clarifies the circumstances in which treating ordinal data with a metric model is likely to yield major discrepancies from using an ordinal model. Moreover, our approach has revealed errors that, to our knowledge, have not been discussed in the literature, such as systematic inversions in group means, cross-over interactions, and trends.

8.4. Frequentist vs Bayesian approaches to ordered-probit models

In a frequentist approach to ordered-probit models, the parameters are estimated by a search algorithm such as gradient ascent that attempts to converge on the maximum likelihood estimate (MLE). Sometimes the algorithms fail to converge or fail to find the global maximum. Even when the MLE is found, *p* values and confidence intervals are estimated by theoretical large-*N* approximation, and are overly-optimistic for small or moderate *N*.

Crucially for our purposes, most packaged frequentist software assumes homogenous variances across groups (as far as we know). This assumption makes it impossible to assess many of the problems we have

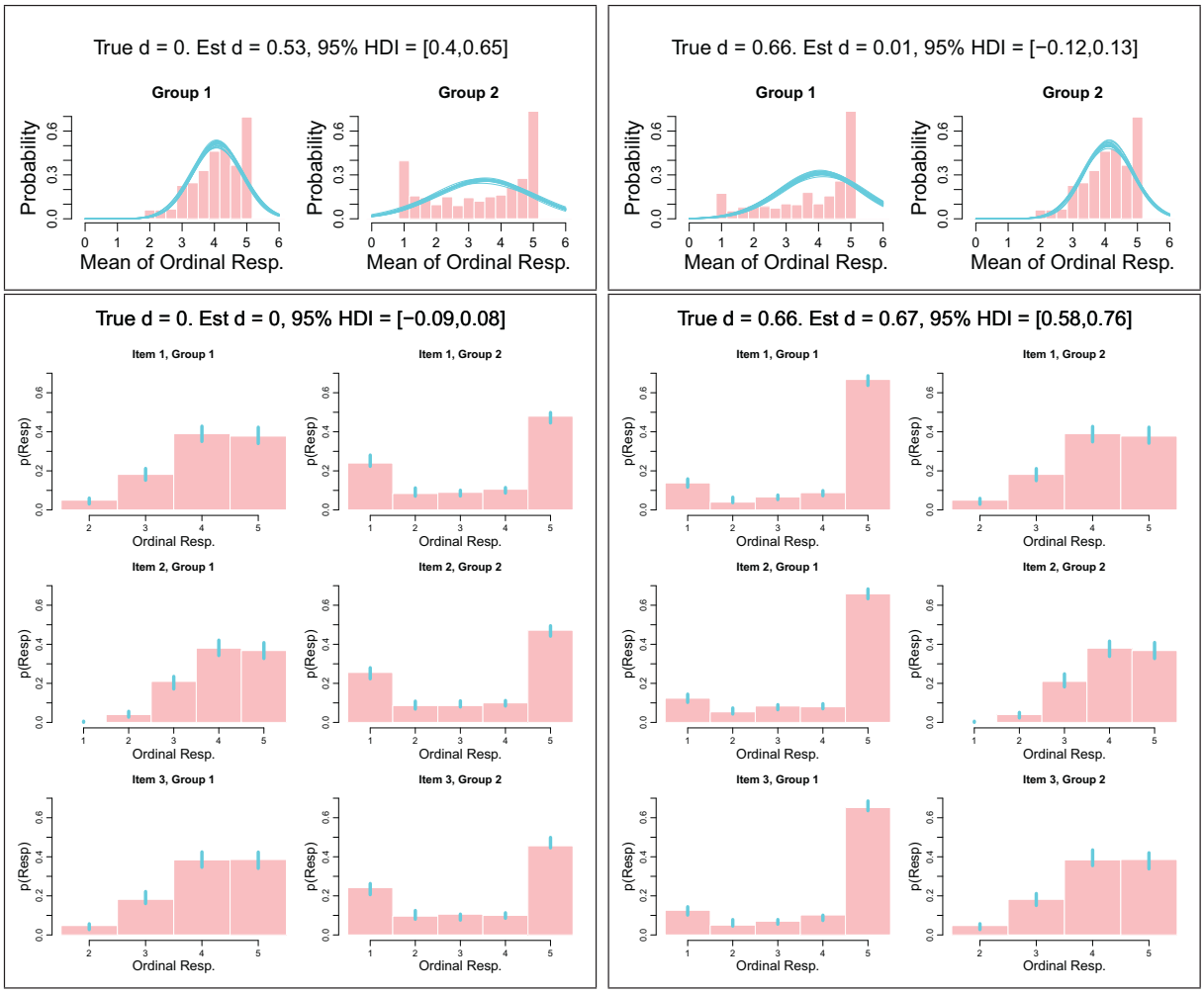

Fig. 10. Ordinal data from three items for two groups, displayed in histograms (same data in upper and lower panels). *Left column:* The generating parameters had $d = 0.0$, corresponding to points Ⓐ and Ⓑ in Fig. 4. The metric model (upper panel) produces a strong false alarm (Type I error), while the ordered-probit model (lower panel) accurately estimates the effect size and mimics the data probabilities. *Right column:* The generating parameters had $d > 0.0$, corresponding to points Ⓑ and Ⓓ in Fig. 4. The metric model (upper panel) completely misses the difference (Type II error), while the ordered-probit model (lower panel) accurately estimates the effect size and mimics the data probabilities.

pointed out in this article, because many (but certainly not all) of the problems arise from unequal variances on the latent variable. In particular, specifying hierarchical structure on heterogeneous variances, as we have as an option in the R script at https://osf.io/53ce9/, may be challenging or impossible in pre-packaged frequentist software.

We prefer Bayesian approaches for a number of reasons. First, software for specifying models is very flexible in popular software such as JAGS (Plummer, 2003) and Stan (Carpenter et al., 2017). In particular, the analysis of movie data presented earlier in this article was carried out using R and JAGS; full details are available in the Appendix and in online supplementary material at https://osf.io/53ce9/. The model has different means and variances for every movie, along with the option for hierarchical structure on the variances across movies. This was done for both the ordered-probit and metric models.

A second reason we prefer Bayesian approaches is that they virtually always converge to accurate values of the parameters and produce accurate values of credible intervals for any *N*. More specifically, Markov chain Monte Carlo (MCMC) methods converge to accurate representations of the posterior distribution, and do so efficiently for typical ordered-probit models and moderately-sized data sets found in psychological research.

A third reason we prefer a Bayesian approach is that it yields a richly informative posterior distribution over the joint parameter space. Every parameter in the model has an exact credible interval. For example, in the ordered-probit model of the movie data, every mean and every standard deviation has an exact credible interval, as do the estimated thresholds. Moreover, exact posterior distributions are obtained for all differences between means across movies, all differences between standard deviations across movies, and differences between thresholds on the latent scale. In trend analyses, slopes, differences of slopes, etc., and their credible intervals can be directly gleaned from the posterior distribution. Essentially anything one wants to know about the parameters can be directly "read off" the posterior distribution.

As mentioned previously, a simple R script for the analysis shown in Figs. 5 through 8 is provided at https://osf.io/53ce9/. The script allows the user to read in any similarly structured data file (i.e., one item measured in multiple groups).

Regardless of whether a frequentist or Bayesian approach is taken, it is worth re-iterating that we are not arguing that an ordered-probit model is the *correct* model of ordinal data. It was important for us to use a probit (i.e., cumulative normal) model because its normal distribution exactly matches the normal-metric model except for the thresholds.

Therefore we could make direct comparisons of ordered-probit models with traditional normal-metric models. When using ordinal models, some people prefer to use a logistic cumulative probability function (logit) instead of a normal cumulative probability function (probit). Logit models will usually yield very similar results to probit models. Bürkner and Vuorre (2018) discussed other ordinal regression models and showed how to implement them in Bayesian software using the brms package (Bürkner, 2017).

8.5. Conclusion

In this article we reported that metric methods are routinely applied to ordinal data in high-impact psychology journals. We showed that treating ordinal data with metric models can produce false alarms (Type I errors) and failures to detect effects (Type II errors). Moreover, metric models of ordinal data can systematically produce *inversions* of effects. We showed these problems run rampant in real data such as movies ratings. We demonstrated the potential for these types errors in other analytical contexts such as interactions and trend analyses. We showed that averaging correlated ordinal items does not solve or even mitigate these problems. Moreover, we argued that these problems can not be diagnosed with certainty from metric models without performing an ordinal analysis. Importantly, we presented graphical explanations for how and when these problems arise: By explicitly displaying means and standard deviations of ordinal values treated as metric, as a function of underlying ordered-probit parameters, it is easy to see the qualitative configurations of parameters that will systematically yield misinterpretations by metric models of ordinal data.

To address these problems, we used an ordered-probit model with parameters estimated through Bayesian analysis. Frequentist approaches to ordered-probit models may also work, but we prefer the flexibility, accuracy, and richness of information provided by Bayesian methods. Because it is impossible to know in advance whether or not treating a particular ordinal data set as metric would produce a different result than treating it as ordinal, we recommend that the default treatment of ordinal data should be with an ordinal model, and we recommend Bayesian estimation as an excellent way to estimate the parameters of such a model.

Appendices

This appendix provides details and expanded discussion of several topics. Its subsections include (i) mathematical details of the models, (ii) expanded discussion of why equal metric standard deviations do not solve the problems of metric models, (iii) examples of errors interpreting interactions, and (iv) examples of errors interpreting trends in regression analysis.

Appendix A. Details of ordered-probit and metric models

For metric models, the probability of the ordinal value ‘y’ is computed by first converting each ordinal label to its corresponding metric value (i.e., ‘1’→ 1.0, ‘2’→ 2.0, ‘3’→ 3.0, and so on) and then describing its probability as the normal density at that metric value. Thus, the probability of ordinal value k is

$$p(y = k|\, \mu^{[g]}, \sigma^{[g]}) = \text{dnorm}\left(\frac{k - \mu^{[g]}}{\sigma^{[g]}}\right) \tag{1}$$

where $\mu^{[g]}$ is the mean of group g, $\sigma^{[g]}$ is the scale (standard deviation) of group g, and dnorm() is the standardized normal density (not the cumulative normal probability). The upper panel of Fig. 1 illustrated an example of Eq. (1). When the metric model is applied to the average of Q ordinal items, each with K ordinal levels, y can take on only the possible values of Q/Q, $(Q + 1)/Q$, $(Q + 2)/Q$, …, KQ/Q. Examples of a distributions over these discrete values were provided in the upper panels of Fig. 10.

For the ordered-probit model, probabilities of ordinal values are computed using the cumulative normal. The probability of ordinal response k is

$$p(y = k|\, \mu, \sigma, \theta_1, ..., \theta_{K-1}) = \Phi\left(\frac{\theta_k - \mu}{\sigma}\right) - \Phi\left(\frac{\theta_{k-1} - \mu}{\sigma}\right) \tag{2}$$

where $\Phi()$ is the standardized cumulative normal function. Eq. (2) says that the probability of ordinal response k is the area under the normal curve between threshold θ_{k-1} and threshold θ_k. For the first level (i.e., $k = 1$) the threshold θ_{k-1} is negative infinity, and for the highest level (i.e., $k = K$) the threshold θ_K is positive infinity. The lower panel of Fig. 1 illustrated the ideas of Eq. (2).

When there is more than one item and more than one group, the probability of ordinal value $y^{[i,g]}$ on item i in group g is specified as:

$$p(y^{[i,g]} = k \mid \mu^{[g]}, \sigma^{[g]}, \theta_1^{[i]}, ..., \theta_{K-1}^{[i]}) = \Phi\left(\frac{\theta_k^{[i]} - \mu^{[g]}}{\sigma^{[g]}}\right) - \Phi\left(\frac{\theta_{k-1}^{[i]} - \mu^{[g]}}{\sigma^{[g]}}\right) \tag{3}$$

Notice that Eq. (3) reduces to Eq. (2) when there is only one item and one group.

The parameter values in the ordered-probit model (i.e., μ, σ, and $\{\theta_k\}$) can trade off and yield identical data probabilities. In particular, a constant could be added to all the thresholds and to the means, but yield the same response probabilities. Independently, all the parameters could be multiplied by a constant but yield the same response probabilities. Therefore two parameter values must be fixed at arbitrary values. We fix the endpoint thresholds at $\theta_1 = 1.5$ and $\theta_{K-1} = K - 0.5$, because then all the parameter values make intuitive sense with respect to the observed values. For instance, if the ordinal scale ranges from ‘1’ to ‘5’ we set $\theta_1 \equiv 1.5$ and $\theta_4 \equiv 4.5$. Then the value of the mean parameter μ and standard deviation parameter σ can be intuitively mapped to the response scale, although this mapping must be done cautiously because it compares a metric latent scale with an ordinal response.

By contrast, the traditional approach to ordered-probit models arbitrarily sets $\sigma = 1$ for all groups and $\mu_1 = 0$, with all thresholds estimated relative to those fixed values of μ and σ (e.g., McKelvey & Zavoina, 1975; Winship & Mare, 1984). But this approach yields parameter values with little intuitive relation to the response values. A transformation for converting the traditional parameterization to our more intuitive parameterization, including a function in the programming language R for converting output from the `polr()` function in the MASS package (Venables & Ripley, 2002), is explained at http://doingbayesiandataanalysis.blogspot.com/2014/11/ordinal-probit-regression-transforming.html with a PDF

version at https://osf.io/fc6zd/. Note, however, the `polr()` function assumes equal variances in all groups, so it cannot be used to reproduce our analyses.

When the predictor is nominal, such as group membership, then the latent mean μ has a different value for each group, indexed by superscript $[g]$. In this case, our models also provide each group with their own standard deviation, $\sigma^{[g]}$. In other words, our models do *not* require homogeneous variances, as is typically assumed by traditional metric models.

Ordered-probit models typically assume that the thresholds (θ_k) are the same across all groups because the thresholds are theoretically linked to the response measure, not to the predictor value. For example, when asked, "How happy are you?" with response options '1' = very unhappy, '2' = mildly unhappy, '3' = neutral, '4' = mildly happy, '5' = very happy," the latent thresholds between ordinal levels are assumed to be implicit in the phrasing of the question, regardless of other aspects of the respondent or situation. In other words, the thresholds are assumed to be part of the measurement procedure, not dependent on the value of the predictor or covariate. This can be technically referred to as a type of measurement invariance.

The sigmoidal curves of Figs. 3 and 4 were computed using the defining formula for expected value:

$$\bar{y} = \sum_{k=1}^{K} k \cdot p(k \mid \mu, \sigma, \theta_1, \ldots, \theta_{K-1}) \tag{4}$$

where $p(k|\mu,\sigma,\theta_1,\ldots,\theta_{K-1})$ is computed using Eq. (2). The plot of ordinal SD as a function of latent mean, in Figs. 11 and 12, was computed by definition as

$$S_y = \left(\sum_{k=1}^{K} (k - \bar{y})^2 \cdot p(k|\mu, \sigma, \theta_1, \ldots, \theta_{K-1}) \right)^{1/2} \tag{5}$$

where $\bar{y}$ is the expected value defined in Eq. (4).

In our Bayesian analyses, we use prior distributions that are broad on all parameters and have minimal influence on the posterior distribution. In the ordered-probit models, the thresholds are given wide normal priors:

$$\theta_k \sim \text{normal}(k + 0.5,\ 2) \quad \text{for} \quad k = 2, \ldots, K - 2 \tag{6}$$

where the second argument in normal(,) is its standard deviation, not its variance or precision. For example, the prior on the threshold between ordinal bins 3 and 4 is centered at 3.5 and given a large standard deviation of 2.0 so that its estimated value can vary widely from 3.5. (The priors specified in Eq. (6) permit inverted thresholds, e.g., $\theta_2 < \theta_1$, but inverted thresholds imply negative probabilities in Eq. (3) and therefore never actually occur.)

The prior on $\mu^{[g]}$ is a broad normal distribution, with a mean set to the midpoint of the ordinal scale and with a standard deviation equal to the number of ordinal categories:

$$\mu^{[g]} \sim \text{normal}((K + 1)/2,\ K) \tag{7}$$

As mentioned before, the second argument in normal(,) is its standard deviation, not its variance or precision. The standard deviation of the prior is very wide so that the estimated means can vary widely.

For the model of movie ratings, movies were indexed by the group superscript $[g]$. The movie standard deviations were described as gamma distributed:

$$\sigma^{[g]} \sim \text{gamma}(\omega_\sigma,\ \sigma_\sigma) \tag{8}$$

where ω_σ is the mode of the gamma distribution and σ_σ is the standard deviation of the gamma distribution. In the default version of the model, the gamma distribution is given a fixed broad prior with $\omega_\sigma = 3.0$ and $\sigma_\sigma = 3.0$.

The R script provided at https://osf.io/53ce9/ also allows an optional hierarchical version of the model, in which the mode and standard deviation were estimated, and given priors:

$$\omega_\sigma \sim \text{gamma}(3,\ 3) \tag{9}$$

$$\sigma_\sigma \sim \text{gamma}(3,\ 3) \tag{10}$$

where the arguments of the gamma distribution refer to its mode and standard deviation. The hierarchical version is useful if the user wants the standard deviations of the different groups to mutually inform each other's estimates, but the hierarchical version can also produce implosive shrinkage if the sample sizes of the groups are small relative to the number of groups.

In the models of movie ratings, the ordered-probit model had 74 parameters: Each of the 36 movies had a mean and standard deviation ($\mu^{[1]}, \ldots, \mu^{[36]}$ and $\sigma^{[1]}, \ldots, \sigma^{[36]}$ from Eq. (3) with a single item), and there were two threshold parameters (θ_2 and θ_3 from Eq. (3) with a single item; recall that the end thresholds were set at $\theta_1 \equiv 1.5$ and $\theta_4 \equiv 4.5$). The metric model had 72 parameters: All the same parameters as the ordered-probit model except the two threshold parameters.

For the model of averaged ordinal items, the prior on $\sigma^{[g]}$ was a broad uniform distribution from 0.01 to ten times the number of response levels:

$$\sigma^{[g]} \sim \text{uniform}(0.01,\ 10 \cdot K) \tag{11}$$

Again, this broad prior distribution is designed to have minimal influence on the posterior distribution.

Bayesian estimation was accomplished through Markov chain Monte Carlo (MCMC) methods to generate a large number of representative parameter value combinations from the posterior distribution. We used the MCMC sampler JAGS (Plummer, 2003), in tandem with the statistical software R and the R package runjags (Denwood, 2016). For more implementation details, see Kruschke (2015). The complete commented R scripts and data files are provided at https://osf.io/53ce9/.

The simulated data in Fig. 2 were created with the `mvrnorm` function from R's `MASS` package using the argument `empirical=TRUE` which creates data that have sample mean and covariance that match the specified generating values. Complete R scripts are available at https://osf.io/53ce9/.

Appendix B. Why equal metric SDs do not avoid problems

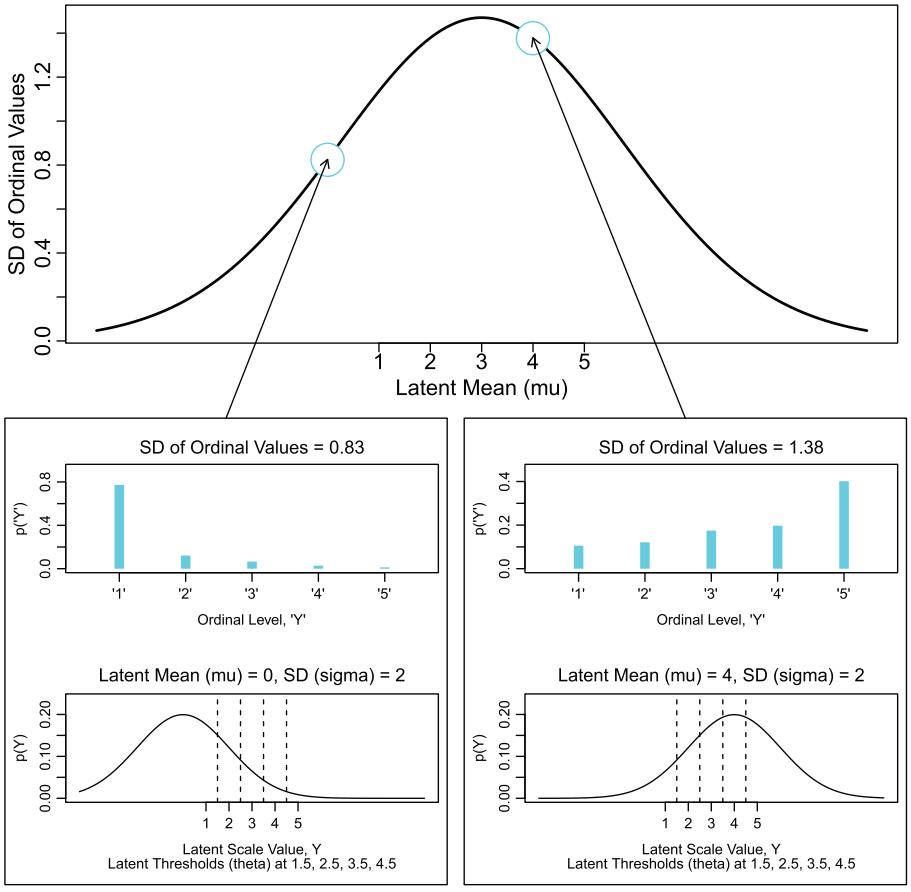

Fig. 11. *Upper panel:* The standard deviation of the ordinal values (from Eq. (5)) is plotted as a function of the latent mean (mu). *Lower panels:* For the two marked points on the sigmoidal curve, lower boxes show the corresponding latent normal distribution and resulting ordinal values. The only difference between the two lowest panels is the mean (mu) of the latent normal distribution; the two lowest panels have the same latent standard deviation (sigma) and the same latent thresholds (theta's).

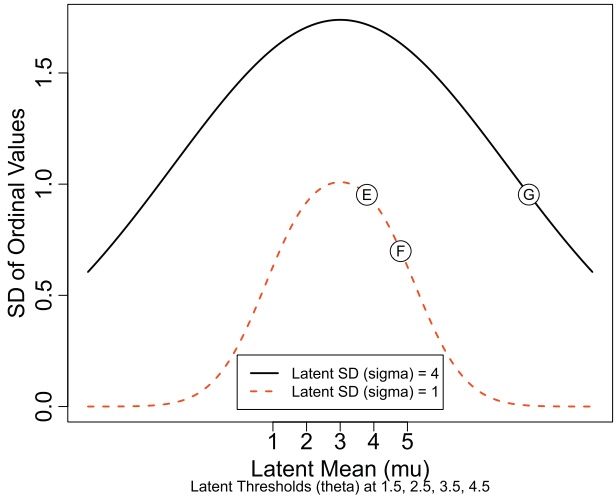

Fig. 12. Ordinal SD as a function of latent mean (mu) and SD (sigma). Letter-labeled points are discussed in the main text.

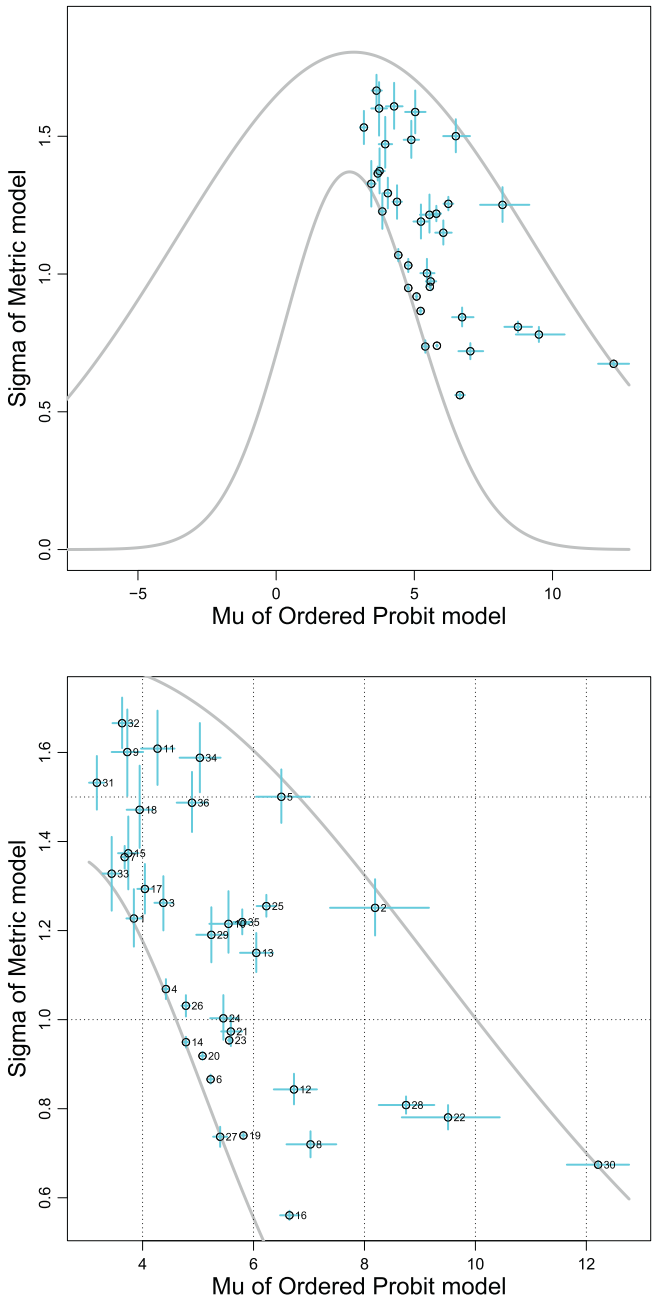

Fig. 13. Movie ratings: Posterior modal sigma values from the metric model are plotted against the posterior modal mu values for the ordered-probit model. Upper panel shows full range as in Fig. 12. Lower panel zooms in with each dot labeled by its case number. Segments intersecting the dots indicate the 95% HDIs of the parameters. Curves are plots of constant ordered-probit sigma as a function of ordered-probit mu for the smallest latent sigma (lower curve) and largest latent sigma (upper curve).

The variances of the ordinal values (treated as metric values) do not reveal the underlying relation of variances in the latent values. Fig. 11 shows the ordinal SD as a function of the latent mean and latent SD. The curve in Fig. 11, computed using Eq. (5), shows that a constant latent SD does *not* produce a constant standard deviation of ordinal values.

Fig. 12 shows the curves of ordinal SD for two values of constant latent sigma. Consider in Fig. 12 the points labeled Ⓔ and Ⓕ. These two groups have the *same* latent σ, as indicated by the fact they fall on the same curve. But their ordinal S_y's are *not* the same. Thus, a difference in ordinal SDs does not imply a difference in latent SDs. Next, consider the points labeled Ⓔ and Ⓖ. These two groups have the same ordinal S_y's, as indicated by the fact they are at the same height along the vertical axis of the plot. But their latent σ's are *not* the same. Thus, equal ordinal SDs do *not* imply equal latent SDs. To reiterate, the SD of the ordinal values does not reflect the latent SD, and therefore we cannot tell from the SD of the ordinal values

whether or not, or to what extent, the mean of the ordinal values reflects the latent mean. Moreover, as was mentioned earlier, groups with equal underlying metric variances such as points Ⓐ and Ⓓ in Fig. 4 tend to have compressed ordinal means and reduced statistical power in the metric model.

Recall the application to movie ratings from the main text. Fig. 13 shows the estimated σ's from the metric model plotted against the estimated μ's from the ordered-probit model, analogous to the layout of Fig. 12. Recall in Fig. 12 the points labeled Ⓔ and Ⓕ, which fall on the same curve of constant σ in the ordered-probit model but have different σ's in the metric model. There are many cases in the movie data of Fig. 13 that show the analogous property. In particular, cases 1, 4, 14, and 27 all have nearly the same ordered-probit σ because they all fall nearly on the curve of constant σ, but those cases have very different σ's in the metric model. In other words, the difference of σ's in the metric model does *not* reflect the nearly equal σ's in the ordered-probit model (which describes the data much better). In Fig. 12 the points labeled Ⓔ and Ⓖ have equal σ's in the metric model but different σ's in the ordered-probit model. The movie data of Fig. 13 show many analogous cases, including 2, 3, and 25, which have nearly the same metric σ but very different ordered-probit σ's (because the cases fall on different curves). In other words, the equality of σ's in the metric model does *not* reflect the very different σ's in the ordered-probit model (which describes the data much better). Fig. 9 in the main text showed the specific comparison of movies 35 and 29 which have nearly the same metric SD, for which the estimated difference in the metric model is not significant but is in the ordered-probit model.

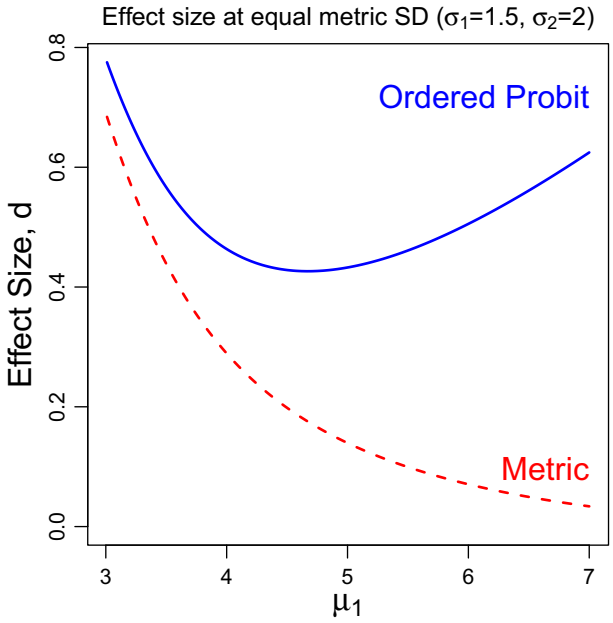

Fig. 14. Effect size when the metric-model standard deviations are equal. Horizontal axis is value of μ_1 (relative to thresholds at $\theta_1 = 1.5$, $\theta_2 = 2.5$, $\theta_3 = 3.5$, and $\theta_4 = 4.5$). The ordered-probit standard deviations are fixed at $\sigma_1 = 1.5$ and $\sigma_2 = 2.0$ for this example. To compute the curves, for any value of μ_1, the value of μ_2 is found that makes the metric SD of group 1 (from Eq. (5)) equal the metric SD of group 2. Then, using that μ_2, the effect sizes are computed and plotted here.

Finally, Fig. 14 shows an example of effect sizes when the metric-model standard deviations are equal. Notice that the effect size in the ordered-probit model is always larger than the effect size in the metric model. Moreover, as the latent mean becomes extreme, the metric-model effect size shrinks to zero while the ordered-probit effect size is large. This qualitative trend persists for any choice of $\sigma_1 \neq \sigma_2$.

Appendix C. Mistakes interpreting interactions

The distortions of estimated means, caused by metric models of ordinal data, can occur in any design. Consider, for example, a design with two independent variables each with two nominal levels, that is, a 2×2 factorial design. For generality, we will assume a "between-subjects" design, in which the data in each cell are independent of the other cells (i.e., the same assumption we made for the two group designs discussed earlier in the article). Suppose that the ordinal data are generated by an ordered-probit model, using the same thresholds in every cell (because we assume the same measurement process in every cell), but with possibly different means and standard deviations in each cell. We now show that the apparent interaction shown by the metric model of the ordinal values can be greatly distorted relative to the true interaction in the latent cell means.

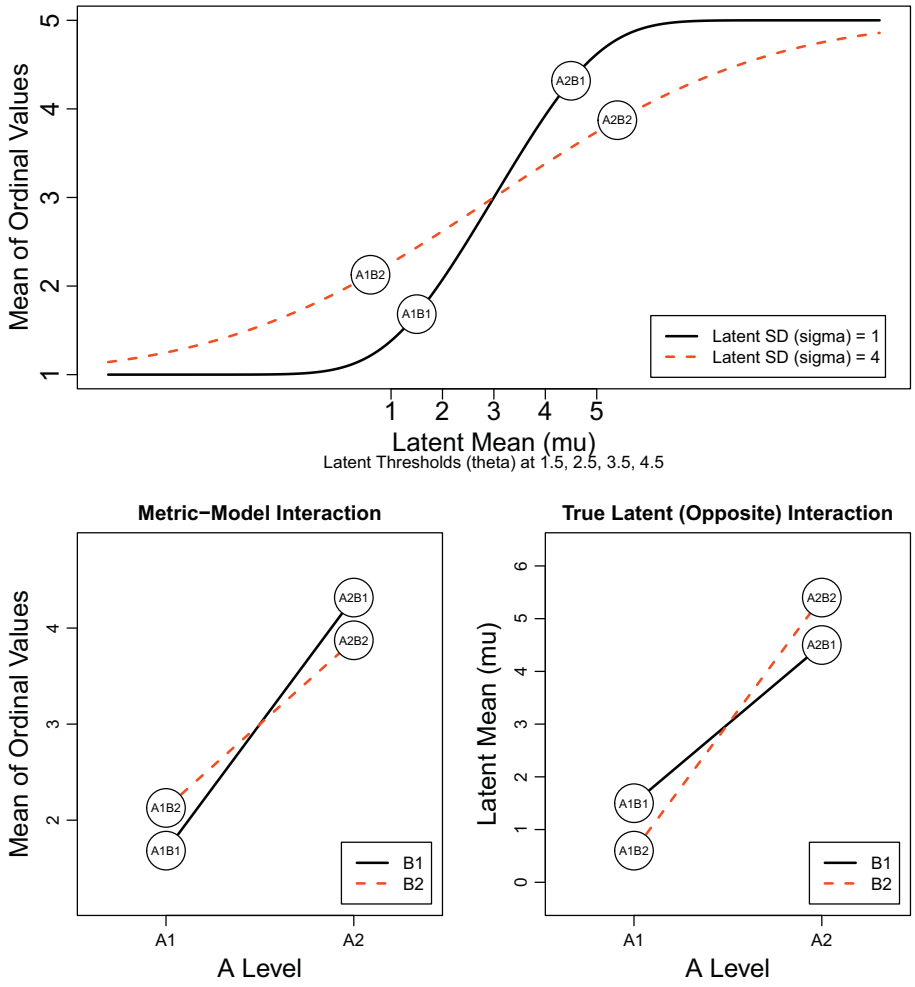

Fig. 15. Example of 2 × 2 factorial design in which the metric model shows a cross-over interaction in one direction, but the true latent values are a cross-over interaction in the *opposite* direction. Upper panel shows the true latent means and standard deviations as points on sigmoid curves in the format of Fig. 4, with the apparent metric mean on the vertical axis. Lower panels show the means plotted as a function of factor levels. Notation: AjBk is level j of factor A and level k of factor B.

Fig. 15 shows a case in which the true latent interaction is a cross-over interaction in one direction, but the interaction shown by the metric model of the ordinal data is a cross-over interaction *in the opposite direction.* In this example, the standard deviation of cells in level B1 is smaller than the standard deviation of cells in level B2, while the true effect of factor A (i.e., the difference between A1 and A2) is smaller at level B1 than at level B2. The upper panel of Fig. 15 shows the true latent means and standard deviations of the four cells, along with the implied metric means. The lower panels of Fig. 15 show the means plotted as a function of factor levels. The lower panels graphically reveal that the cross-over interaction shown by the metric model is in the opposite direction of true cross-over interaction in the latent means.

It would be straightforward to generate large-*N* random samples from the ordered-probit model in Fig. 15 and demonstrate that, indeed, the metric model shows a cross-over interaction opposite the true cross-over interaction of the ordered-probit model. The demonstration would be trivial because large-*N* random samples will accurately reflect their generating parameter values. The demonstration would show, again, that the ordered-probit model more accurately describes the data than the metric model, as was demonstrated in numerous previous examples. We refrain from including the numerical simulation to save space.

The example of Fig. 15 relies on the inversion of means pointed out previously as cases © and ⓓ in Fig. 4, applied twice. The qualitative inversion of the cross-over does not rely on the specific parameter values chosen in the figure; obviously the exact values can be “slid around” while maintaining the ordinal relations of the means.

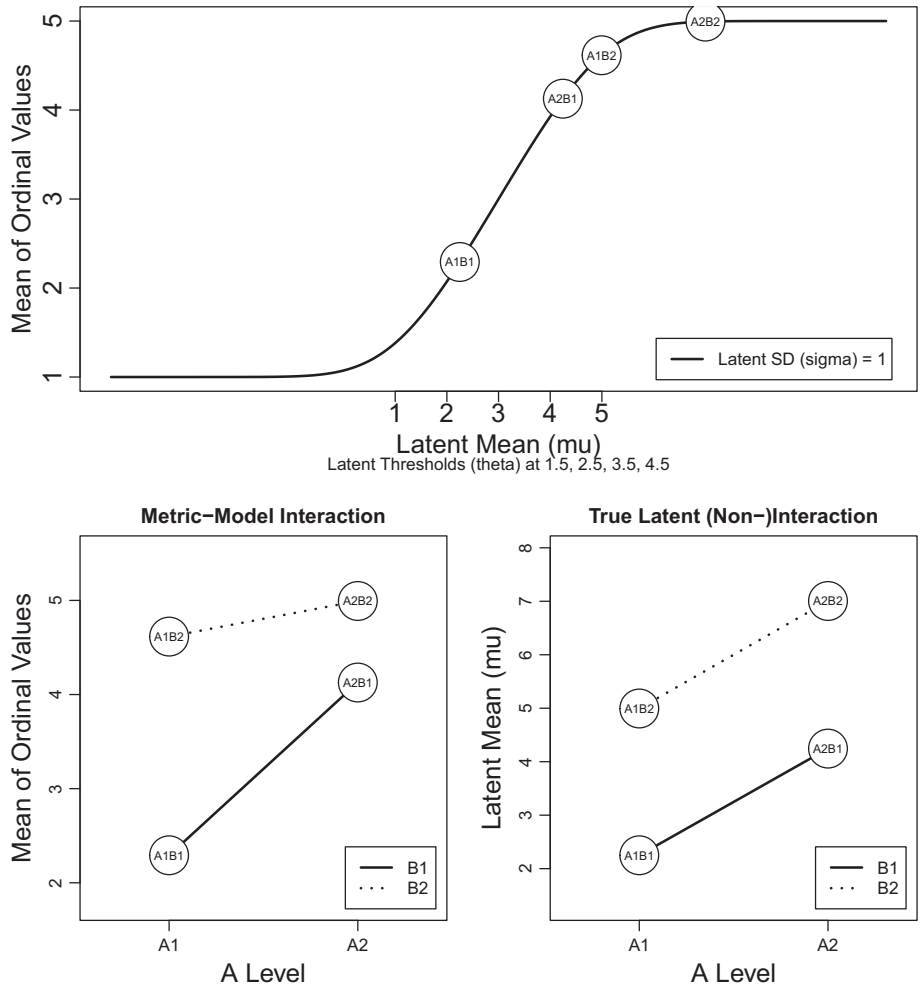

Fig. 16. Example of 2 × 2 factorial design in which the metric model shows an interaction, but the true latent values have *zero* interaction. In this example all groups have identical latent standard deviation. Upper panel shows the true latent means and standard deviations as points on a sigmoid curve in the format of Fig. 4, with the apparent metric mean on the vertical axis. Lower panels show the means plotted as a function of factor levels. Notation: AjBk is level j of factor A and level k of factor B.

Fig. 16 shows another example of a mis-interpreted interaction, in this case a false-alarm (Type I error) that arises even when the underlying standard deviations are identical in all cells. The upper panel of Fig. 16 shows the latent means and standard deviation along with the metric-model means of the ordinal data. Because of the sigmoidal constriction of the metric means, the equal spacing of latent means is transformed into unequal spacing of metric-model means. The lower panels of Fig. 16 show the means plotted by factor levels, and graphically show the interaction apparent in the metric-model means despite the lack of interaction (i.e., parallel lines) in the latent means.

Of course it is straightforward to construct another example in which there is zero interaction in the metric-model means but non-zero interaction in the latent means, which would be a Type II error by the metric model. Such a case is constructed by selecting equally spaced intervals on the metric-model axis (i.e., the vertical axis of the upper panel of Fig. 15), finding those heights on the sigmoidal curves, and projecting to the corresponding latent means, which will not be equally spaced.

In summary, when applied to interactions, metric-model means of ordinal data can produce Type I errors (false alarms), Type II errors (misses), and even directional inversions of cross-over interactions. To what extent do these situations arise in real data? The answer depends utterly on the distribution of data in whatever domain is being investigated, and we cannot know how the natural world has chosen to distribute its ordinal data. But we do know that the ordered-probit model will be a better description of ordinal data than the metric model, and therefore an ordinal model of ordinal data is always a better approach than an analogous metric model.

Appendix D. Linear regression and trend analysis

As the reader may have intuited by now, the non-linear (sigmoidal) mapping from latent mean to ordinal mean will also distort estimates of regression coefficients, whether linear or non-linear. Consider linear regression, in which the latent mean is a linear function of the predictor variable, x. The latent mean is mapped to an ordinal response, and the metric mean of the ordinal responses is a sigmoidal function of the latent mean (as has been shown repeatedly, for example in Fig. 4). In other words, the true linear latent trend becomes a sigmoidal trend on the ordinal scale. But if the ordinal responses are modeled directly with a linear function, the linear function is being erroneously superimposed on a sigmoidal trend. The problem is exacerbated with unequal bin widths in the ordered-probit model. Therefore, metric linear regression and trend analysis applied directly to the ordinal values will mis-estimate the regression coefficients, mistakenly inferring differences or non-differences in slopes across conditions, and mistakenly inferring presence or absence of non-linear trends.

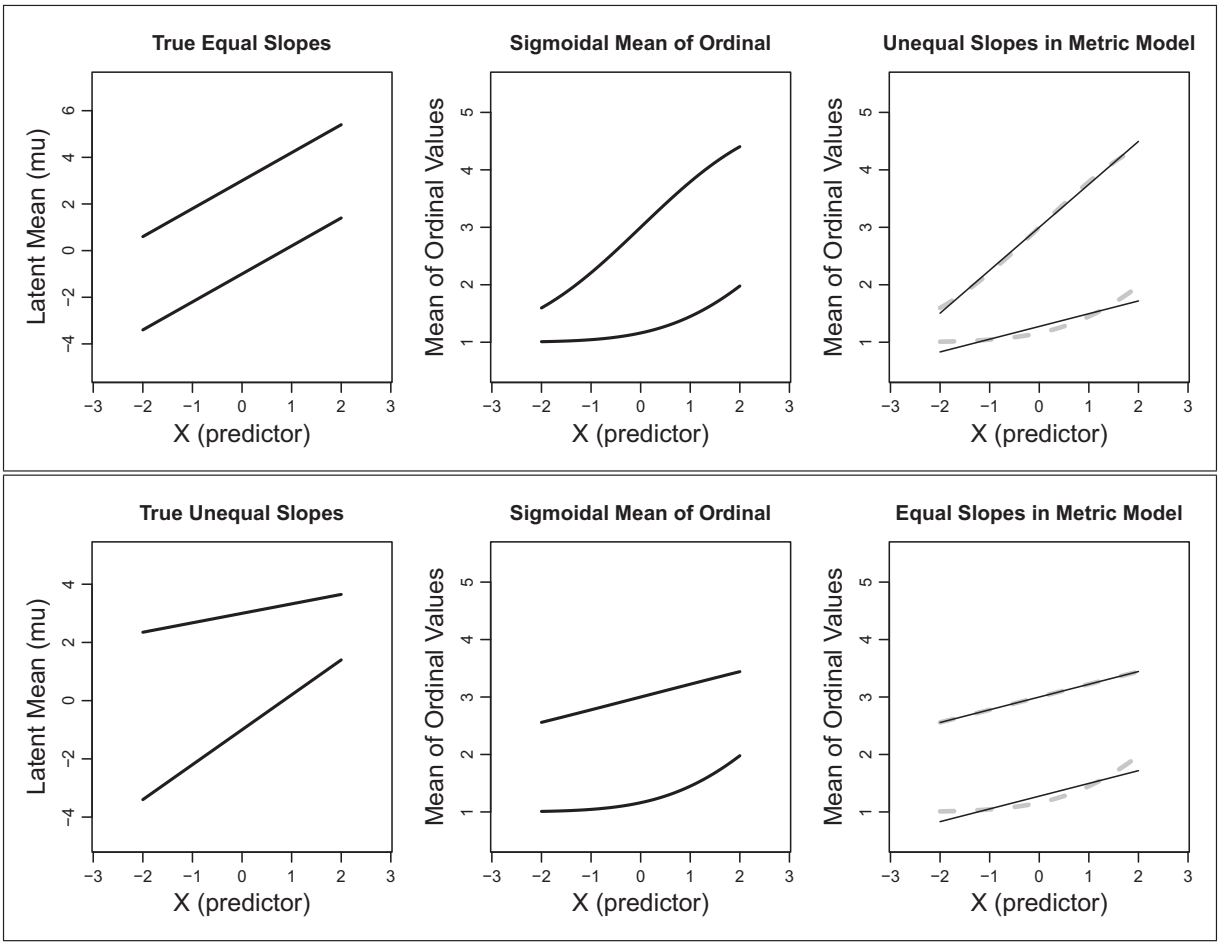

Fig. 17. Examples of simple linear regression for two groups of data. For both rows, the horizontal axis is the predictor variable, x, on an arbitrary scale. The two curves in each panel correspond to the trends in two groups. The ordered-probit model is assumed to have thresholds at 1.5, 2.5, 3.5, and 4.5, with $\sigma = 2.0$. *Upper row:* True latent slopes are equal (left panel) but metric linear regression model of ordinal value shows unequal slopes (right panel). *Lower row:* True latent slopes are not equal (left panel) but metric linear regression model of ordinal values shows equal slopes (right panel).

Consider a hypothetical example in which subjective happiness, rated on an ordinal response scale, is modeled as a linear function of annual income, for people from two groups (e.g., two different climates, two political parties, or two pre-questionnaire mood-induction treatments). Suppose that the true influence of the groups is only a difference in intercepts (i.e., overall happiness), but the slopes of the two groups are identical (i.e., the increase in happiness as a function of income is the same in both groups). Fig. 17 illustrates this case in the figure 's upper row. The left panel shows the true underlying linear trends in the two groups, with *equal* slopes. The middle panel shows the sigmoidal transformation of these linear trends into the ordinal response scale. The right panel shows a metric linear regression model fit to the two sigmoidal trends, in which the slopes are distinctly *not* equal. The lower row of Fig. 17 shows a complementary situation in which the true latent trends do *not* have equal slopes but the metric linear regression model of the ordinal values shows *equal* slopes in the two groups.

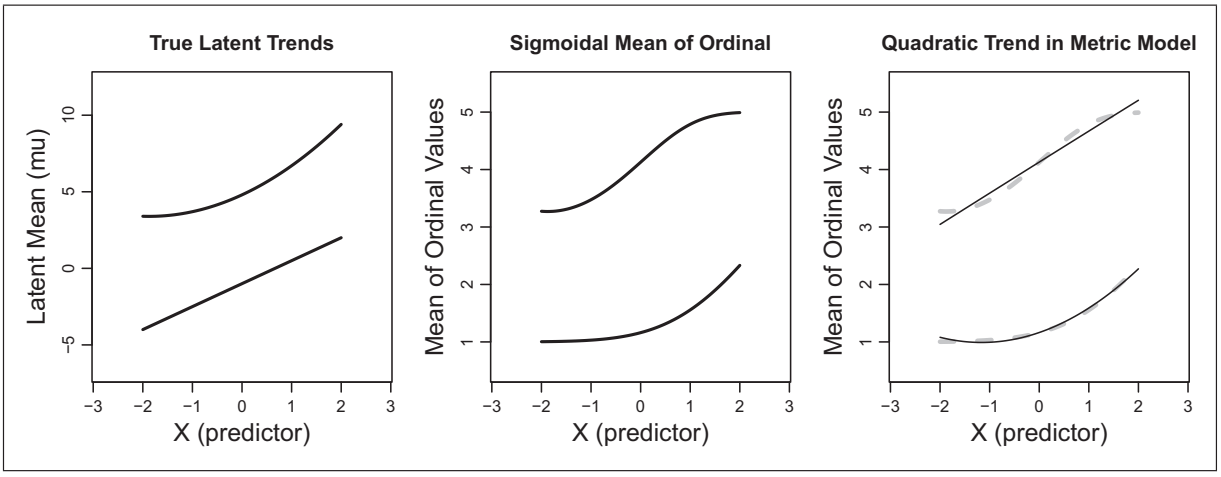

Fig. 18. Examples of quadratic trend analysis for two groups of data. The horizontal axis is the predictor variable, x, on an arbitrary scale. The two curves in each

panel correspond to the quadratic trends in two groups. The left panel shows that the true trend in the lower group has zero quadratic trend while the upper group has a positive quadratic trend. The middle panel shows the sigmoidal transformation of the trend to the ordinal scale. The right panel shows quadratic trend analysis treating the ordinal values as if they were metric. The metric model erroneously shows zero quadratic trend in the upper group and positive quadratic trend in the lower group. (The ordered-probit model is assumed to have thresholds at 1.5, 2.5, 3.5, and 4.5, with $\sigma = 2.0$.)

Non-linear trend analysis is also distorted. In particular, quadratic trends can be mistakenly confabulated or suppressed. Fig. 18 shows an example in which the true latent trend (left panel) has zero quadratic trend for the lower group and positive quadratic trend for the upper group, but the metric quadratic trend model of the ordinal values produces the opposite result: zero quadratic trend for the upper group and positive quadratic trend for the lower group.

References

Albert, J. H., & Chib, S. (1997). Bayesian methods for cumulative, sequential and two-step ordinal data regression models. *Technical report, bowling green state university*.

Babakus, E., Ferguson, C., & Jöreskog, K. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violation of measurement scale and distributional assumptions. *Journal of Marketing Research, 29*(May), 222–228.

Becker, W. E., & Kennedy, P. E. (1992). A graphical exposition of the ordered probit. *Source: Econometric Theory, 8*(8), 127–131. https://doi.org/10.1017/S0266466600010781.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*(1), 49–64. https://doi.org/10.1037/h0041412.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C., & Vuorre, M. (2018). *Ordinal regression models in psychological research: A tutorial. **, **(**):*https://osf.io/cu8jv/.

Carifio, J., & Perla, R. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 3*(3), 106–116.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42*(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), https://doi.org/10.18637/jss.v076.i01.

Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education, 35*(4), 31–35. https://doi.org/10.5032/jae.1994.04031.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Denwood, M. (2016). runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software, 71*(9), 1–25. https://doi.org/10.18637/jss.v071.i09.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*(2), 309–326.

Feldman, M. P., & Audretsch, D. B. (1999). Innovation in cities: Science-based diversity, specialization and localized competition. *European Economic Review, 43*, 409–429. https://doi.org/10.1016/S0014-2921(98)00047-6.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods, 9*(4), 466.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237–288. https://doi.org/10.3102/00346543042003237.

Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson correlation against violations of assumptions. *Perceptual and Motor Skills, 43*(3f), 1319–1334. https://doi.org/10.2466/pms.1976.43.3f.1319.

Heeren, T., & D'Agostino, R. (1987). Robustness of the two independent samples t test when applied to ordinal scaled data. *Statistics in Medicine, 6*(1), 79–90.

Hsu, T. C., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the F-test. *American Educational Research Journal, 6*(4), 515–527. https://doi.org/10.3102/00028312006004515.

Hui, M., & Bateson, J. E. G. (1991). Perceived control and the effects of crowding and consumer choice on the service experience. *Journal of Consumer Research, 18*(2), 174–184.

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.

Kruschke, J. K. (2015). *Doing Bayesian data analysis, second edition: A tutorial with R, JAGS, and Stan*. Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1.

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4.

Likert, R. (1932). *A technique for the measurement of attitudes*. New York: Columbia University Press.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514–534.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103–120. https://doi.org/10.1080/0022250X.1975.9989847.

Mooijaart, A. (1983). Two kinds of factor analysis for ordered categorical variables. *Multivariate Behavioral Research, 18*(4), 423–441. https://doi.org/10.1207/s15327906mbr1804_5.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(38), 171–189. https://doi.org/10.1111/j.2044-8317.1992.tb00975.x.

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: Theory and Practice, 15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y.

O'Brien, R. M. (1979). The use of Pearson's r with ordinal data. *American Sociological Review, 44*(5), 851–857. https://doi.org/10.2307/2094532.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika, 23*(1), 114–133.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76–105.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677.

Stevens, S. S. (1955). On the averaging of data. *Science, 121*(3135), 113–116. https://doi.org/10.1126/science.121.3135.113.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Vickers, A. J. (1999). Comparison of an ordinal and a continuous outcome measure of muscle soreness. *International Journal of Technology Assessment in Health Care, 15*(04), 709–716.

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review, 49*(4), 512–525. https://doi.org/10.2307/2095465.